Évaluation de la concordance dans les études d'imagerie diagnostique : une étude de la qualité des données publiées

Jules Tyaniu Zhang-Yin

AVERTISSEMENT

Cette thèse d'exercice est le fruit d'un travail approuvé par le jury de soutenance et réalisé dans le but d'obtenir le diplôme d'Etat de docteur en médecine. Ce document est mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt toute poursuite pénale.

Code de la Propriété Intellectuelle. Articles L 122.4

Code de la Propriété Intellectuelle. Articles L 335.2-L 335.10

UNIVERSITÉ PARIS DESCARTES

Faculté de Médecine PARIS DESCARTES

Année 2015 N° 64

THÈSE

POUR LE DIPLÔME D'ÉTAT

DE

DOCTEUR EN MÉDECINE

Évaluation de la concordance dans les études d'imagerie diagnostique : une étude de la qualité des données publiées

Présentée et soutenue publiquement

le 18 juin 2015

Par

Jules Tianyu ZHANG-YIN

Né le 11 mai 1986 à Guilin (Chine)

Dirigée par M. Le Professeur Philippe Chaumet-Riffaud, PU-PH

Jury :

M. Le Professeur Pierre Weinmann, PU-PH …….. Président

M. Le Professeur Alain Prigent, PU-PH

Mme Le Docteur Claire Vaylet de Labriolle, MCU-PH

Table des matières

1.INTRODUCTION .....

1.1 PROBLEMATIQUE ET OBJECTIFS..... 7
1.2 NOTION DE CONCORDANCE.....10
1.3 PRINCIPAUX TESTS STATISTIQUES APPLICABLES DANS LA PROBLEMATIQUE DE LA CONCORDANCE.....12

2. MATERIELS ET METHODES.....

2.1 SELECTION DES REVUES..... 17
2.2 SELECTION DES ARTICLES..... 18
2.3 ELABORATION ET JUSTIFICATION DES ITEMS..... 20
2.4 METHODOLOGIE POUR LA CONCORDANCE SENIOR/JUNIOR..... 26
2.5 STATISTIQUES..... 27

3.RESULTATS.....

3.1 DESCRIPTION GENERALE DES ARTICLES..... 28
3.2 ITEMS REMARQUABLES..... 28
3.3 ITEMS GLOBAUX..... 36
3.4 CONCORDANCE SENIOR/JUNIOR..... 41

4.DISCUSSION.....

4.1 REPRESENTATIVITE ET EXTRAPOLABILITE DES ARTICLES..... 43
4.2 HYPOTHESES SUR LES ITEMS REMARQUABLES..... 44
4.3 HYPOTHESE SUR LES DISCORDANCES (SENIOR/JUNIOR)..... 45
4.4 PORTEE DE L'ETUDE..... 46

5. CONCLUSION..... 47

6. FIGURES (ANNEXE 1)..... 48

7. TABLEAUX (ANNEXE 2)..... 50

8. LEXIQUE (ANNEXE 3)..... 57

9. BIBLIOGRAPHIE..... 60

Remerciements

A Monsieur le Professeur Pierre WEINMANN que j'ai l'honneur d'avoir comme Président du Jury. Je tenais à vous exprimer toute ma gratitude pour la formation (notamment en cardiologie nucléaire) durant mon passage dans votre service, votre efficacité, votre très grande réactivité. A l'avenir, j'aurais toujours grand plaisir à discuter de la cuisine asiatique avec vous.

A Monsieur le Professeur Philippe CHAUMET-RIFFAUD. C'est un très grand honneur pour moi d'avoir eu la chance de travailler avec vous durant mon internat et de vous avoir comme Directeur de Thèse. Vous m'avez appris le goût de l'excellence, de la réflexion, de la rigueur, de la discussion et de la perfection en méthodologie.

A Monsieur le Professeur Alain PRIGENT d'avoir accepté de faire partie du Jury et d'avoir beaucoup contribué à ce travail. En effet, sans votre participation très active et motivée et ce même parfois pendant vos jours de congés, je n'aurais certainement pas pu finir à temps. Merci pour votre disponibilité, votre gentillesse et votre écoute tout au long de ce parcours.

A Mme La Docteur Claire VAYLET de LABRIOLLE d'avoir accepté de faire partie du Jury et de votre soutien généreux quant à mon futur poste d'AHU. Je serai très content de pouvoir collaborer avec vous dans un futur proche et espère pouvoir profiter de votre expérience dans le domaine de la pédiatrie.

A Monsieur Vincent CHEKIB. Merci pour ta disponibilité et ton écoute malgré un emploi de temps très chargé. Sans ton apport dans le domaine de recherche bibliographique, ce travail n'aurait jamais pu voir le jour.

A tous ceux qui m'ont accompagné tout le long de mon internat et de ma formation:

A l'équipe de médecine nucléaire de l'hôpital René Huguenin où j'ai fait mes premiers pas d'interne. J'ai beaucoup appris et j'en retiendrais les leçons tout au long de ma carrière.

A l'équipe de radiothérapie de l'hôpital René Huguenin et l'équipe de médecine nucléaire de l'HEGP qui ont beaucoup contribué à ma formation.

A l'équipe de médecine nucléaire de l'hôpital Hôtel-Dieu (Karim, Xavier et Nadine) et de l'institut Curie (Slavomir et Florent) où j'ai été bien accueilli lors de mes quelques passages.

A l'équipe de médecine palliative de l'hôpital Necker (Marcel-Louis, Roger et Perrine) où j'ai beaucoup appris sur le plan humain. Spéciale dédicace à Marcel-Louis, ta gentillesse et ta connaissance en philosophie m'ont illuminé.

A l'équipe de médecine nucléaire du Centre cardiologique du nord (Bernard, David, Mohamed et Mathieu) où j'ai reçu une excellente formation dans une très bonne ambiance. Je suis fier d'être le 1e interne du service et cela m'a permis de découvrir l'exercice dans le privé.

A l'équipe de radiologie de l'hôpital militaire de Percy (notamment Julien et Mr Baccialone). Dans ce stage j'ai beaucoup appris et acquis la confiance nécessaire pour gérer seul certaines situations pas toujours faciles. Merci pour votre accompagnement et votre formation. Je me suis senti parfois comme chez moi ici.

A l'équipe de médecine nucléaire de l'hôpital militaire de Val-de-Grâce (notamment Eric, Mathieu et Denis). J'ai toujours voulu faire un stage dans ce service dès le début de mon internat. J'ai enfin réalisé ce rêve et hélas j'ai assisté à l'annonce de la fermeture de ce très bel hôpital rempli d'histoire et qui est un symbole de l'excellence. J'ai reçu une très bonne formation et ai constaté cette volonté de toujours bien faire le travail malgré les circonstances difficiles. Je suis fier d'avoir fait partie de ce service et je le considère comme un héritage.

A l'équipe de médecine nucléaire de l'hôpital Princesse Grâce à Monaco pour votre chaleureux accueil. Je suis sûr que je vais y passer un très bon semestre et je suis content de retrouver Pr Faraggi qui m'a donné l'envie de m'engager dans cette belle spécialité qu'est la médecine nucléaire.

A l’équipe de médecine nucléaire de l’hôpital Saint-Antoine pour votre soutien à mon futur poste dans le service. Je suis de ravi de pouvoir collaborer avec vous.

A mes anciens co-internes (Loc, Caroline, Ophélie et Aurélien) avec qui j’ai passé des moments agréables.

A mes amis de la fac (notamment Sylvain et Dris) avec qui j’ai partagé le banc pendant 5 ans dans notre belle et réputée faculté de Paris Descartes. A nos sorties et soirées sans oublier des moments de travail intense pour réviser le concours.

A mes co-internes parisiens de médecine nucléaire (Bénédicte, Céline, Ophélie, Marc, Eve, Laura…). Nous avons la chance d’avoir une très bonne ambiance et certains sont devenus mes meilleurs amis.

A mes amis de Saclay (notamment Erwan, Yassine, Marie et Philippe). Je suis fier de notre promo exceptionnellement unie, ce qui nous permet de bien nous engager dans les missions d’ANAIMEN.

A mes parents sans qui je ne serais pas là aujourd'hui, merci pour votre soutien en toutes circonstances et je crois que tous les mots sont superflus quand il s’agit de vos mérites.

A ma grand-mère grâce à qui j’ai acquis cette affinité pour la médecine (et aussi un peu pour la médecine militaire).

A France et Pierre pour votre accueil dans la famille et votre gentillesse.

A Clotilde, merci pour ton amour, ta présence et ton soutien.

Et enfin, à cette belle spécialité de médecine nucléaire !

1. INTRODUCTION

1.1 PROBLEMATIQUE ET OBJETIFS

Les progrès médicaux et scientifiques sont fondés sur l'acquisition de nouvelles connaissances qui doivent être robustes et fiables pour servir de support à la définition des Bonnes Pratiques médicales.

Dans le domaine de la recherche biomédicale, nous sommes confrontés depuis de nombreuses années à l'accumulation sans précédent de nouvelles données en raison, entre autres, de la multiplication des vecteurs de diffusion de résultats scientifiques.

Malheureusement, il a été constaté que la majorité de ces découvertes ne tiendra pas l'épreuve du temps [1]. En effet, les résultats ne sont souvent pas reproduits entre les différentes études, ce qui met en doute la validité des données et leur interprétation.

D'une manière plus générale, cette « crise » de la reproductibilité dans la recherche fondamentale et en expérimentation préclinique pourrait être le résultat du non-respect des bonnes pratiques scientifiques notamment méthodologiques et d'une volonté ou d'un besoin « excessif » de publier à tout prix (publish or perish). [2]

Heureusement, au cours de ces dernières années, la prise de conscience des faiblesses qui existent dans notre système actuel de la recherche fondamentale et préclinique s'est accrue.

En effet, plusieurs études [3-5] ont souligné le fait qu'un grand nombre de recherches précliniques étaient incapables de reproduire des résultats présentés, et ceci même pour des publications sorties dans des journaux ayant un facteur d'impact très élevé.

Les estimations du taux de résultats non reproductibles varient de 75 % à 90 % dans ces observations empiriques. Curieusement, ces estimations correspondent globalement bien aux estimations de 85 % qui représentent la proportion de la recherche biomédicale qui sont estimées gaspillées par certaines équipes [6-7].

Ce très mauvais taux de reproductibilité ne se limite pas seulement aux travaux précliniques mais semble aussi être un fait dans le domaine de la recherche biomédicale.

Afin d'illustrer mon propos, quelques articles vous seront présentés.

Begley et Ellis en 2012 ont étudié 53 études de type essai clinique qui avaient été publiés dans le journal « Oncology » ; 90% des résultats de ces études ne sont pas reproductibles [4].

Concernant les études animales dans le domaine des pathologies neurologiques, Tsilidis et al. (2013) ont conclu qu'un grand nombre de résultats étaient certes significatifs mais qu'ils présentaient des biais manifestes de sélection. Ces articles ont été publiés dans « Neurological studies » [8].

Perrin a repris en 2014 les résultats de 100 études sur le traitement expérimental de sclérose latérale amyotrophique chez les souris in vivo : aucun résultat n'a pu être reproduit par la suite [9].

Ioannidis JP et al. ont rapporté que dans 16 études sur 18 traitant d'analyses d'expression génique par microarray, les résultats n'étaient pas reproductibles [10].

Dans la littérature médicale, en particulier lors des dernières années, nous avons identifié quelques publications non reproductibles qui avaient fait l'objet de plusieurs centaines de citations.

Pour les responsables de la santé publique, le principal défi actuel est d'obtenir des résultats valides, car les financements destinées à la recherche sont limités et donc précieux. Il faut ainsi veiller à leur utilisation le plus efficace possible dans le cadre des appels d'offre. Dans le domaine préclinique, il est déplorable que les résultats de beaucoup de recherches ne puissent être reproduits, y compris par les équipes à l'origine des publications princeps. Il s'agit donc d'un sujet important en termes de retombées potentielles pour la santé publique, la communauté scientifique, les organismes de recherche (Inserm, ANR…) et les décideurs politiques.

Le sujet est complexe en raison de la diversité des acteurs impliqués, et des systèmes actuels de financement et d'évaluation de la recherche ; il n'existe à l'évidence pas un seul responsable, ni une solution unique. Nous nous proposons d'analyser les différentes situations pouvant conduire à ces problèmes de reproductibilité en recherche biomédicale fondamentale et préclinique. En comprenant les spécificités de chaque situation, nous envisagerons des options de solutions qui pourraient aider à améliorer la qualité de la recherche et sa reproductibilité.

L’interprétation et la prise en compte du résultat positif d’une recherche s’apparente à la notion d’une valeur prédictive positive. Plusieurs méthodologistes ont suggéré [11-13] que le taux élevé de non reproductibilité de travaux pouvait être la conséquence d’un excès de confiance dans des résultats de tests significatifs (généralement avec une valeur de p fixée à 0,05), sans se soucier de la plausibilité des hypothèses testées et de la pertinence de la méthodologie employée dans l’étude. Si le pourcentage de plausibilité des hypothèses testées est faible, le taux de faux positifs sera élevé.

La valeur de p (ou niveau de significativité) permet de mesurer la force (ou la robustesse) du choix de rejeter l'hypothèse nulle. Plus la valeur de p est faible, moins l'hypothèse nulle est probable. Le seuil de significativité fixé à 0,05 résulte d’un choix arbitraire et n’a pas de justification formelle. Beaucoup d’auteurs ont insisté sur le fait qu’il serait plus adéquat de fixer le seuil à 0,01 pour assurer la solidité des conclusions.

Cette problématique de reproductibilité existe également dans le domaine de la recherche en imagerie qui a ses spécificités. La place de l’imagerie est aujourd’hui majeure dans les progrès médicaux car de nombreuses procédures d’imagerie orientent le diagnostic (dépistage, diagnostic, stadification….) et la prise en charge thérapeutique (évaluation de la réponse thérapeutique, actes radioguidés, imagerie interventionnelle…). La pertinence de ces procédures d'imagerie peut varier en fonction de la situation clinique. Le transfert de nouvelles techniques d’imagerie en pratique clinique, ce qui signifie une substitution possible à la technique de référence, requiert au préalable d’établir la concordance entre la nouvelle méthode et la ou les procédures considérées comme référence.

Une méthode est désignée comme étant la méthode de référence si elle permet de refléter la réalité de manière la plus fidèle sur des bases physiopathologiques, bien que non dénuée d’imprécision intrinsèque et inévitable.

En imagerie médicale, de nombreuses études ont eu pour objectif de valider des nouvelles techniques plus facilement utilisables, ou ayant des meilleures performances diagnostiques (sensibilité, spécificité, exactitude ou précision) que les méthodes de référence.

Une méthodologie rigoureuse et clairement définie est nécessaire pour planifier ce type d'étude et des tests statistiques adéquats sont également indispensables pour évaluer l'accord entre les deux méthodes.

Or dans la littérature médicale, il n'existe pas encore d'étude portant sur cette problématique.

En effet, jusqu'au présent, les travaux méthodologiques portent essentiellement sur la performance diagnostique des examens, se basant sur la publication « Towards Complete and Accurate Reporting of Studies of Diagnostic Accuracy: The STARD Initiative » [14] parue en 2003 qui propose 25 items permettant de contrôler la qualité méthodologique de ce type d'études. Mais aucun travail similaire portant sur les études traitant la concordance et la reproductibilité n'est disponible.

Le premier objectif de ce travail a été d'évaluer la qualité méthodologique et la pertinence des analyses statistiques portant sur un échantillon représentatif de travaux d'imagerie médicale à visée diagnostique publiés durant les 10 dernières années dans des journaux internationaux de bon niveau. Nous avons proposé des critères de qualité pour évaluer chacune de ces études (sous la forme d'une check-list). Le travail a également comparé les résultats obtenus par un médecin junior (un interne) et deux médecins séniors expérimentés.

1.2 NOTION DE CONCORDANCE

La performance d'une nouvelle méthode de mesure s'évalue par sa concordance (agreement en anglais) avec la méthode de référence. Il est important de rappeler la notion de la concordance.

Définition de la concordance :

Il s'agit de la capacité d'une méthode de mesure à fournir une valeur d'un paramètre quantitatif aussi proche que possible de celle obtenue avec une autre méthode de mesure considérée par les hommes de l'art comme la référence [15-18]. La concordance peut aussi s'apprécier pour la cotation de variables ordinales ou binaires (cas classique de la comparaison de jugements fournis par plusieurs lecteurs).

Très souvent on observe dans la pratique courante, les termes « fiabilité » et « concordance » sont souvent utilisés de façon interchangeable. Cependant, les deux concepts sont conceptuellement distincts.

La concordance et la fiabilité répondent à deux questions différentes:

1. " Quel est l'accord entre les mesures répétées ? " Il s'agit de l'erreur intrinsèque qui mesure et évalue la dispersion des scores de mesure répétés.
2. " Quelle est l'exactitude de la mesure ? " Il s'agit de la véracité des mesures qui reflète bien ou non la maladie.

Définition de la corrélation :

On peut définir la corrélation comme : l'existence d'une interdépendance entre des variables quantitatives.

Il est nécessaire de bien distinguer la notion de la corrélation et celle de la concordance, car l'utilisation de tests de corrélation statistique n'est pas pertinente pour l'évaluation de la concordance [19, 20]. Certains auteurs ont utilisé la régression linéaire entre les valeurs données par la méthode de référence et celles obtenues par la nouvelle méthode ; malheureusement ils ont souvent employé la valeur du coefficient de corrélation pour apprécier la concordance. Or cette procédure ne permet pas de tester la concordance.

En effet, la recherche statistique d'une corrélation est à employer uniquement quand on s'intéresse à des mesures de deux variables distinctes (poids et taille d'un sujet par exemple). Par contre, la mesure de la même variable par deux méthodes distinctes donnera deux séries de valeurs qui seront toujours corrélées (deux méthodes certes différentes mais mesurant le même paramètre au même instant et chez un même patient) ; ces mesures ne peuvent pas être indépendantes donc elles seront toujours corrélées. Un exemple extrême serait celui de la mesure de la taille de sujets avec deux toises électroniques très précises (très bonne reproductibilité) mais dont l'une donne une valeur deux fois plus faible que l'autre. Dans ce cas, il y aura une corrélation statistique remarquable et égale à 1 alors que la concordance est évidemment catastrophique entre ces deux groupes de données.

Nous proposons un lexique (annexe 3) qui s'est inspiré d'une recommandation récente en 2010 de FDA pour la définition des différents termes.

1.3 PRINCIPAUX TESTS STATISTIQUES APPLICABLES DANS LA PROBLEMATIQUE DE LA CONCORDANCE

Alors que le problème de l'évaluation de la concordance entre les données qualitatives ou quantitatives est assez fréquent dans la pratique clinique, il n'y a que peu de tests disponibles en pratique.

Il convient de distinguer deux types de tests : ceux adaptés aux données qualitatives et aux données qualitatives.

Tests pour données qualitatives (pour des variables binaire ou ordinale)

Le test kappa est un test non paramétrique qui a été proposé par J Cohen en 1960, et qui mesure l'accord entre différents observateurs qui ont coté qualitativement des observations (le plus souvent, il s'agit d'un classement qualitatif en catégories). Le coefficient kappa de Cohen: (15, 18, 21-22) a été développé pour étudier l'agrément entre deux observateurs. L'accord entre les jugements reflète la conformité des informations données par les deux observateurs sur un objet identique (jugements appariés sur une variable de même nature).

La « concordance » signifie la proportion de sujets pour lesquels il y a accord entre les observateurs. On évalue cette concordance grâce au calcul d'un coefficient nommé « kappa », qui est d'autant plus proche de 1 que la concordance est bonne.

Le calcul du coefficient kappa se réalise à l'aide des paramètres suivants :

- la concordance observée Po (somme des proportions diagonales du tableau de contingence)
- la concordance calculée Pc (concordance attendue sous l'hypothèse d'indépendance).

Et sa formule de calcul est la suivante:

Kappa = (Po- Pc)/(1-Pc)

A noter qu'il est nécessaire de connaître les limites de l'utilisation de ce coefficient kappa. En premier lieu, le nombre de catégories influe sur la valeur du coefficient et la réduction du nombre de catégories (par regroupement de classes) augmente la valeur du kappa. Dans certains cas, on est amené à estimer que certaines discordances de classification entre deux juges sont plus graves que d'autres ; il a été proposé dans ces cas d'employer un kappa

pondéré ce qui nécessite d'associer au tableau de contingence des données une matrice des poids associés aux désaccords. Or le choix de ces poids est très subjectif. Enfin les proportions des observations dans chacune des cases du tableau de contingence doivent être équilibrées. En imagerie, le coefficient kappa est dépendant de la prévalence du signe recherché (Feinstein et Cichetti).

Il existe une interprétation de kappa généralement admise qui est pourtant subjective et donc discutable. [15]

Tableau

Interprétation du coefficient Kappa proposée par
Landis et Koch

Coefficient Kappa	**Estimation du degré de concordance**
0,8 à 1	Excellent
0,6 à 0,8	Bon
0,4 à 0,6	Moyen
0,2 à 0,4	Faible
0 à 0,2	Négligeable
0	Mauvais

Fleiss a étendu ce kappa aux mesures de concordance impliquant plus de deux observateurs en proposant une version modifiée en 1981.

Tests pour données quantitatives

Le graphique de Bland-Altman

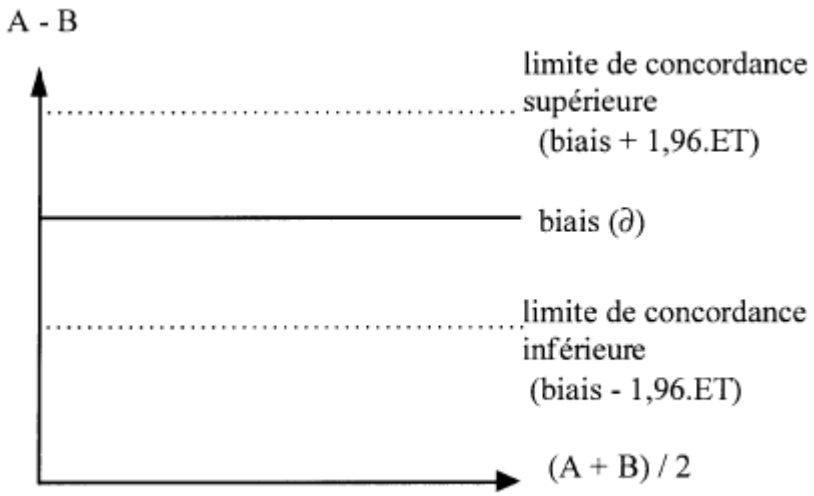
En 1986, Bland et Altman ont proposé une méthode permettant de représenter graphiquement la concordance [20]. Il ne s'agit pas à proprement parler d'un test statistique et il n'y aura pas de valeur de p fournie. L'interprétation de cette représentation doit se faire avec les cliniciens concernés car il faudra leur demander quelle différence de mesure entre deux dispositifs est acceptable pour leur pratique quotidienne.

Cette méthode permet de comparer les moyennes de mesures à leurs différences. Pour cela, il faut donc reporter sur un graphique les points représentant les résultats des deux mesures ; ces points ont pour coordonnées en abscisse la moyenne de chaque couple de valeurs (valeur nouvelle méthode + valeur méthode de référence divisée par 2) et en ordonnée la différence entre les deux valeurs obtenues.

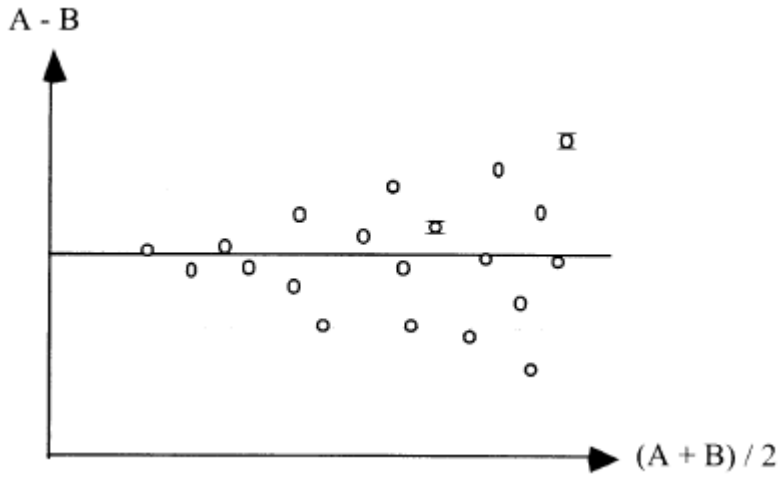
Dans cette représentation, la moyenne des différences est représentée sur ce graphe par une droite parallèle à l'axe des abscisses. De plus, deux droites pointillées sont tracées qui vont délimiter l'intervalle qui devrait comprendre 95 % des valeurs expérimentales (si distribution gaussienne de part et d'autre du biais absolu). Ces limites d'agréments sont donc équidistantes de la droite du « biais absolu moyen » et leur ordonnée est égale au biais absolu moyen +/-1,96 écart-type du biais absolu. On emploie très souvent le terme de biais pour la moyenne des différences mais c'est un abus de langage ; en fait cette moyenne des différences ne représente le biais que si l'un des deux systèmes de mesure est le gold standard et qu'il y a un nombre suffisant d'observations. Dans le cas général, il s'agit plus d'une différence systématique entre les résultats de deux méthodes de mesure qu'un biais au sens strict de la statistique.

Il est possible de calculer des intervalles de confiance sur la valeur de la différence moyenne et sur les limites d'agrément. Naturellement, plus l'échantillon d'observations sera grand, plus les intervalles de confiance seront étroits.

Exemples de graphique Bland Altman



Ex 1 : le biais absolu moyen et les zones de 95 % des valeurs du biais absolu sans différences des moyennes



Ex 2 : Différences (petit rond) proportionnelles à l'amplitude des moyennes

Rappelons que les mesures de corrélation ne doivent pas être employées car il ne s'agit que de mesures de l'association. Le test de corrélation (test de Pearson) est fondée sur l'hypothèse que la pente de la droite de corrélation est différente de 0 ; or pour la concordance, l'objectif est de d'analyser l'écart entre la droite de corrélation des observations et la bissectrice du plan qui indique la corrélation parfaite.

ICC (intraclass correlation coefficient) : coefficient de corrélation intraclasse

Le coefficient de corrélation intra classe est basé sur l'analyse de la variance (ANOVA) et notamment dans le cadre des modèles à effet aléatoire. Il existe plusieurs types de mesures de l'ICC [23].

L'ICC est utilisé pour évaluer la cohérence (au sens homogénéité ou uniformité), ou la conformité de mesures faites par des méthodes différentes de la même quantité sur un objet physique. Il peut s'agir, par exemple de la quantification d'un score de calcification sur un TDM par des radiologues en l'absence de standard de vérité. Il s'agit alors de s'assurer de la similarité des mesures entre les praticiens pour lesquels il existe à la fois une composante de variabilité inter-observateur mais aussi intra-observateur. La variabilité inter-observateur traduit les différences systématiques entre les cotateurs, ce qui donne systématiquement des

scores plus bas. La variabilité intra-observateur correspond aux déviations du score d'un observateur donné pour un sujet donné.

Le coefficient intra-classe permet d'étudier la "corrélation" (le degré d'association) entre une variable nominale (ou catégorielle) et une variable quantitative. En analyse de la variance, le carré de ce coefficient permet d'évaluer la proportion de variance "expliquée" par un facteur ou par une interaction entre facteurs (coefficient dit d'intensité de l'effet). Sur le plan pratique, le calcul de cet indice dépend des caractéristiques de la situation à laquelle on est confronté.

Il y a six grandes méthodes décrites pour le calcul des coefficients ICC et implémentées dans les logiciels standards. Le choix de ces formules ICC repose sur l'objectif du travail, le plan méthodologique de l'étude et du nombre et procédé de recueil des mesures,

Nous suivrons la terminologie employée entre autres par le logiciel SPSS™ pour classifier les modèles. Dans le modèle 1, chaque objet de la mesure est évalué par un ensemble différent d'observateurs tirés au sort (schéma très rarement employé dans les études d'imagerie). Pour le modèle 2, chaque objet de la mesure est évalué par le même ensemble d'observateurs mais ces observateurs ont été tirés au sort parmi l'ensemble des observateurs possibles (cette méthode qui utilise un échantillon « représentatif » de la population des observateurs est très souvent employée dans les études d'imagerie). Enfin dans le modèle 3, chaque objet de la mesure est évalué par le même ensemble d'observateurs mais ces observateurs représentent la totalité des observateurs pertinents (cette méthode ne correspond pas à la majorité de situations rencontrées dans les étude. d'imagerie). Après avoir regardé la façon dont est constitué le panel de cotateurs, il faut s'intéresser à la méthode prise en compte des mesures. Il sera possible d'étudier la fiabilité sur la base d'une mesure unique, ou de la moyenne de deux mesures par les cotateurs, voire de la moyenne de 3 mesures ou plus. En pratique dans le domaine de l'imagerie, la valeur de n est fixée à 1 dans la majorité des études. Du point de vue mathématique, un coefficient de corrélation intra-classe est un rapport entre la variance d'intérêt et la variance totale. La variance provient de la méthode de mesure ou d'autres termes. Si ICC= 0, cela signifie que la variance totale ne vient que de la différence entre les méthodes de mesure ou entre les observateurs. En termes d'interprétation, une valeur d'ICC supérieure à 0,75 est considérée comme très bonne et moyenne entre 0,4 et 0,75. La méthode d'ANOVA employée est déterminée en fonction du design de l'étude :

- un échantillon de sujets et un groupe d'expérimentateurs déterminés imposent un modèle dit modèle mixte à deux facteurs (Two-way Mixed) ;
- un échantillon de sujets et un groupe d'expérimentateurs de type aléatoires imposent un modèle dit modèle à effets aléatoires à deux facteurs (Two-way Random). Ceci correspond par exemple à la question de savoir si la détection d'une lésion par un radiologue dépend du radiologue et donc de son expérience ;
- un échantillon de sujets et un groupe d'expérimentateurs de type aléatoires mais dans lequel le choix de l'investigateur pour chaque sujet est aléatoire. Dans ce cas, il faut utiliser un modèle dit modèle à effets aléatoires à 1 facteur (One-way Random).

2. MATERIELS ET METHODES

2.1 SELECTION DES REVUES

La sélection des articles a été faite conformément aux recommandations pour ce type d'analyse et avec l'aide d'un spécialiste en bibliométrie et analyse documentaire. Nous avons utilisé le site http://impactfactor.weebly.com/ (spécialisé dans les journaux médicaux) qui fournit le classement de revues en fonction de leur impact factor et de leur discipline.

La spécialité « radiology » a été sélectionnée pour la recherche de revues. En effet dans le site http://impactfactor.weebly.com/, les revues de radiologie et de médecine nucléaire sont regroupées dans la rubrique « radiology ».

Nous avons retenu les revues ayant les «impact factors» les plus élevés.

Nous avons volontairement écarté les revues ayant un lien manifeste avec une autre discipline ou une technique donnée (par exemple la revue *Neuroimage* ou *Circulation cardiovascular imaging* ou encore *Ultrasound obstetricgynecoloy* malgré leur impact factor plus élevés) afin de préserver l'aspect neutre lors de la sélection de revues.

Par souci d'équité, il a été sélectionné quatre revues de médecine nucléaire et quatre revues de radiologie.

Au final les revues retenues sont :

- En médecine nucléaire :

Journal of Nuclear Medicine (Impact factor : 5,774)

European journal of Nuclear Medicine and Molecular Imaging (Impact factor : 5,114)

Molecular Imaging (Impact factor : 3,408)

Clinical Nuclear Medicine (Impact factor : 2,955)

- En radiologie:

Radiology (Impact factor : 6,339)

Investigative radiology (Impact factor : 5, 46)

European Radiology (Impact factor : 3,55)

American journal of Roentgenoloy (Impact factor : 2,90)

2.2 SELECTION DES ARTICLES (cf. Fig. 1)

Les articles de ces revues ont été analysés de façon consécutive sur une période de dix ans débutant le 1e juillet 2004 et se terminant le 1e juillet 2014. Lors de la première étape de recherche, les mots clé « agreement et comparison » ont été utilisés. Nous avons volontairement limité le nombre de mot clé, à deux en l'occurrence, pour permettre une recherche exhaustive et ainsi augmenter la validité de l'étude.

Les articles publiés en version papier ou en version électronique (Epub) ont été accepté. Grâce à la base de données MEDLINE (via PubMed), nous avons identifié après cette première étape (fig.2) 372 articles répondant aux critères de recherche initiaux. Ce travail a été validé par un spécialiste en recherche bibliographique (Vincent Chekib). La deuxième étape de recherche a consisté en une lecture du résumé des articles afin d'éliminer les articles non pertinents sur la base des informations synthétiques.

Ont été exclus les articles :

- concernant l'expérimentation animale,

- concernant des travaux sur des sujets décédés,
- rapportant des comparaisons de logiciels de traitement des images (par exemple des algorithmes de reconstruction),
- traitant des comparaisons de différents produits de contraste ou des radio-pharmaceutiques
- sur les traitements / études dosimétriques,
- comparant l'aspect sémiologie de l'imagerie versus les données moléculaire/génétique/physiopathologique ou épidémiologique
- revues à visée didactique,
- concernant un échantillon de sujets trop petit (effectif inférieur à cinq),
- concernant plus d'une indication/pathologie.

Il restait 192 articles après cette seconde étape de la sélection. Nous avons effectué une lecture intégrale du texte afin de constituer le lot final des articles. Cette troisième et dernière étape a consisté à s'assurer que les articles présélectionnés traitaient réellement de la problématique de concordance.

Ainsi ont été éliminés les articles qui :

- abordaient uniquement les évaluations des performances diagnostiques,
- et ceux qui ne mentionnaient pas dans le chapitre « matériel et méthodes » de test relevant de l'évaluation de la concordance (absence de mention pour le test kappa ou la mesure de l'ICC ou d'utilisation de graphes type Bland-Altman dans la partie statistique).

Au final, 123 articles ont été identifiés selon ce processus. Nous avons décidé d'exploiter le contenu de 80 articles en raison du souhait d'équilibrer le nombre d'articles provenant de revues de radiologie et le nombre d'articles tirés de revues de médecine nucléaire (le nombre de revues de radiologie est beaucoup plus important que celui de médecine nucléaire). Cette démarche est justifiée car un effectif de 80 articles constitue déjà un échantillon représentatif.

2.3 ÉLABORATION ET JUSTIFICATION DES ITEMS

Nous avons décidé d'analyser ces articles avec une grille de lecture comportant des items spécifiques aux différentes parties. Pour la création de ces items, nous nous sommes inspirés notamment des propositions élaborées dans l'article *Guidelines for Reporting Reliability and Agreement Studies (GRRAS)* du *Journal of Clinical Epidemiology* 64 (2011) [25]. Ce document traitait de la problématique de la concordance, mais sans être spécifique aux études dans le domaine de l'imagerie. L'objectif des auteurs était d'élaborer une trame (guideline) qui pourrait être utile pour les chercheurs, les auteurs, les reviewers et les rédacteurs de revues.

Les différents items qui ont été retenus sont (tableau 1):

Pour la partie « TITRE ET RESUME »

Item N.1 : Identifier si les mots clés «agreement» ou «comparison» sont présents dans le titre

Item N.2 : Identifier si les mots clés «agreement » ou «comparison» sont présents dans le résumé

Item N.3 : Identifier si les mots clés «agreement» ou «comparison» sont dans les mots clés de l'article

Justification : les bases de données bibliographiques (notamment la Medline via Pubmed) et internet sont devenus les ressources primaires pour la recherche dans une démarche de "Evience Based Medicine".

Pour utiliser efficacement ces bases de données, il est nécessaire de pouvoir identifier de manière rapide et claire les études traitant le sujet de la concordance. Pour bien les discriminer des autres articles, nous recommandons de bien spécifier (au moins 1) les mots clés "agreement" ou "comparison" à la fois dans le titre, le résumé et les mots-clés fournis par le texte.

En effet, comme il a été rappelé dans l'introduction: il existe souvent une confusion entre la notion de corrélation et de concordance. L'absence de standardisation des mots clés conduit

non seulement à une perte de temps dans la recherche, mais aussi au risque de confondre les études qui ne sont pas conçues pour la problématique.

Pour la partie « INTRODUCTION »

Item N.4 : Les objectifs de l'étude portent-ils bien sur la concordance?

Justification : certaines études prétendent (par leur titre ou résumé) porter sur la problématique de la concordance mais en réalité elles ont pour objectifs primaires d'évaluer la performance diagnostique de la technique (sensibilité, spécificité, valeur prédictive positive et valeur prédictive négative).

Item N.5 : Les méthodes de mesure sont-elles explicitement et précisément nommées (pas seulement en abréviation) et décrites?

Justification : le degré de fiabilité/concordance est lié aux propriétés des techniques/instruments. Il peut exister différentes versions des appareils ou de techniques de mesure, ce qui pourrait entraîner des compréhensions différentes voire des confusions chez les évaluateurs en fonction de la formulation utilisée.

Item N.6 : L'affection est-elle bien définie?

Justification : il est nécessaire de bien définir l'indication dans laquelle cette concordance est validée puisque la substitution d'une technique s'inscrit dans le cadre bien défini d'une pathologie donnée.

Pour la partie « MATERIELS ET METHODES »

Concernant la population étudiée :

Item N.7 : Pour la population étudiée : les méthodes d'inclusion sont-elles bien décrites?

Justification : la sélection de la population est toujours importante puisque les dispositifs de mesure ou de diagnostic sont souvent conçus pour l'exploration d'une population spécifique. Les caractéristiques des sujets conditionnent l'interprétation de la concordance, parce que les résultats obtenus sont étroitement liés à cette population spécifique.

Item N.8 : Pour la population étudiée : le nombre de sujets de la population étudiée est-il bien défini ?

Justification: comme dans toute étude, il est nécessaire d'avoir un nombre de sujets inclus suffisant pour constituer l'échantillon permettant de réaliser des études statistiques pertinentes.

Item N.9 : Pour la population étudiée : avons-nous l'information précise sur la distribution de l'âge (moyenne et médiane) dans la population étudiée ?

Justification : il s'agit d'un item classique de la caractéristique de la population.

Item N.10 : Avons-nous l'information précise sur le sexe ratio?

Justification : il s'agit également d'un item classique de la caractéristique de la population.

Concernant les évaluateurs (synonymes observateurs, lecteurs, juges…) :

Item N 11 : Le nombre des évaluateurs est-il bien précisé ?

Justification : il est nécessaire de connaître leur nombre (et les modalités de leur sélection) pour appliquer le test statistique le plus adéquat.

Item N 12 : L'expérience des évaluateurs est-elle bien décrite ?

Justification : des années de formation ou d'expérience sont nécessaires au cursus des imageurs. Il est donc évident que la qualité d'une interprétation dépend fortement de l'expérience de l'évaluateur. En conséquence, le résultat d'une mesure de concordance peut varier de manière importante en fonction de l'expérience des évaluateurs.

Mentionner «expérimenté» sans précision est insuffisant n'apporte peu d'information permettant de juger le niveau d'expérience.

Item N 13 : Les évaluateurs sont-ils issus du même centre ?

Justification : le fait que les évaluateurs soient issus du même centre ou non influence l'interprétation et donc la concordance. Il faudrait le mentionner explicitement. En effet, les personnes issues du même centre ont le plus souvent suivi une formation et un apprentissage similaire d'où une technique d'interprétation similaire (meilleure reproductibilité intra-centre mais avec risque de biais lié à un effet centre).

Item N 14 : Le recueil des données se fait-il de manière prospective ou rétrospective ?

Justification : une étude rétrospective présente souvent plus de biais d'une étude prospective.

Item N 15 : Existe-il une justification du calcul de taille d'échantillon ?

Justification : comme dans toute étude, la connaissance préalable et la justification de la taille de l'échantillon sont nécessaires.

Concernant le déroulement de l'examen :

Item N 16 : existe-il une description précise du déroulement de l'examen ?

Justification : cela permet de vérifier qu'un protocole de réalisation des examens avait bien été rédigé (modalités précises de la technique) ; ceci facilite la détection des écarts et déviations au protocole.

Item N 17 : L'activité administrée est-elle bien précisée (pour les examens de médecine nucléaire) ?

Justification : il s'agit d'un point fondamental pour garantir la reproductibilité du résultat dans le domaine de l'imagerie fonctionnelle. L'activité injectée devrait être conforme aux recommandations actualisées dans une indication donnée ou justifiée dans le protocole de la recherche en cas d'utilisation hors AMM.

Item N 18 : le délai entre l'administration du radio-pharmaceutique et l'acquisition des images est-il bien précisé (pour les examens de médecine nucléaire) ?

Justification : il s'agit d'un autre point fondamental pour garantir la reproductibilité du résultat dans le domaine de l'imagerie fonctionnelle. Tout non-respect du délai entraînera des conséquences sur la validité du résultat. Là aussi, ce délai devrait être conforme aux recommandations actualisées dans une indication donnée ou aux données de la littérature.

Item N 19 : le protocole d'acquisition des images est-il bien précisé ?

Justification : il est nécessaire de vérifier si le design du protocole est bien adapté pour répondre à la question clinique.

Item N 20 : L'interprétation est-elle en insu et bien décrite?

Justification : Comme dans les essais thérapeutiques, l'insu est fondamental pour garantir la validité externe. D'autant plus que dans une étude de concordance, les relecteurs ne devraient pas être influencés par la connaissance des autres données pour interpréter les données d'imagerie.

L'article doit indiquer si l'insu a été maintenu total (pas d'information sur les données cliniques, biologiques ou sur d'autres données d'imagerie) ou partiel (une partie seulement). Il s'agit d'un point essentiel.

Item N 21 : Existe-il une description de la procédure de définition des classes de résultats (binaire, ordinal, classe, variable quantitativeetc..) ?

Justification : la connaissance de la procédure de classement conditionne le choix de test statistique. Comme cela a été explicité dans la partie 1.3, il convient d'utiliser la méthode Bland-Altman ou les coefficients ICC pour les variables quantitatives, tandis que le coefficient kappa de Cohen est le plus adapté pour des variables qualitatives.

Item N.22 existe-t-il une comparaison par rapport au gold standard ?

Justification : il est toujours important de se référer au gold standard lorsqu'il est techniquement possible et éthiquement acceptable puisqu'un gold standard est censé d'être plus proche de la vérité.

Item N.23 Existe-t-il une description des tests statistiques utilisés ?

Item N. 24 Ces tests statistiques sont-ils pertinents?

Justification: Il existe plusieurs approches statistiques qui peuvent être utilisées dans la mesure de la fiabilité et de la concordance. Comme ils ont été souvent développés dans différents contextes ou disciplines, il n'existe pas une approche unique pouvant être considérée comme la norme. Chaque méthode est également fondée sur des hypothèses en fonction de la nature des données (nominales, ordinales, quantitative en continu), de la technique d'échantillonnage (choisi au hasard, consécutif et de la gestion des biais potentiels). En conséquence, il n'est pas possible d'être trop normatif concernant la « meilleure » méthode statistique, il faudrait essayer de moduler en fonction de l'objectif et de la conception de l'étude.

Pour les parties « RESULTATS ET DISCUSSION »

Item N.25 est-ce que le nombre de sujets analysés est identique à celui donné pour le nombre de sujets inclus?

Item N.26 Existe-t-il un tableau récapitulant les motifs de sortie de l'essai ?

Justification: le perdu de vue est toujours un risque de biais de sélection qu'il faut vérifier et contrôler. Les auteurs doivent fournir des justifications si le nombre de sujets analysés est différent de celui donné pour le nombre de sujets inclus.

Item N.27 Existe-t-il une description des caractéristiques démographiques des sujets inclus ?

Justification: la connaissance des caractéristiques démographiques permettra de juger de la validité externe du résultat et de l'extrapolation ou non des résultats à d'autres populations.

Item N.28 Existe-t-il une ou des déviation(s) au protocole?

Justification: le respect strict du protocole conditionne la qualité méthodologique de l'étude ; il faudrait mentionner explicitement si le protocole a bien été respecté et justifier toute déviation au protocole.

Item N.29 Existe-t-il une pertinence clinique de l'interprétation ?

Justification: En matière d'évaluation de la concordance, il n'y a pas de chiffre « significatif» ou non (coefficient de Kappa ou de l'ICC par exemples) car il s'agit d'une mesure de l'accord entre observateurs (et non pas d'un test statistique), qui doit être analysé en fonction du contexte.

S'il s'agit d'un examen qui a une valeur décisive sur le pronostic vital ou fonctionnel du patient, on tolérera beaucoup moins une concordance même bonne et on exigera une concordance quasiment parfaite. En effet, il est normal d'être strict vis-à-vis du choix de l'examen de substitution. Ainsi les auteurs devraient dans la discussion aborder cet aspect.

2.4 METHODOLOGIE POUR LA CONCORDANCE SENIOR/JUNIOR

Deux PU-PH (seniors) (professeur des universités – praticien hospitalier) du département biophysique/médecine nucléaire des hôpitaux universitaires Paris Sud site de Bicêtre et un interne (junior) de DES de médecine nucléaire ont réalisé l'analyse de l'ensemble des articles sélectionnés (80 au total).

Cette analyse a été réalisée de façon indépendante et a duré 8 mois (entre avril 2014 et janvier 2015). Nous avons commencé par plusieurs réunions au cours desquelles nous avons échangé nos points de vue concernant la définition des différents items. Après relecture d'environ 60 articles (vers novembre 2014), nous nous sommes entretenus afin de s'assurer qu'il n'existait pas de discordance majeure dans la compréhension conduisant à une interprétation biaisée.

Au cours de l'analyse, chacun de nous s'est muni d'une grille de lecture et nous avons répondu aux différents items.

2.5 ANALYSES STATISTIQUES

Compte tenu des contraintes de temps, il a été décidé de mettre dans un premier temps l'accent sur le versant qualitatif de l'analyse. Ainsi nous nous sommes concentrés sur une analyse descriptive des données.

Ont été calculés :

- le pourcentage de chaque modalité pour chaque item de la grille de lecture
- le taux de concordance entre l'évaluation du junior et l'évaluation des seniors

Certains items se sont révélés particulièrement intéressants en raison de défaut d'information souvent constaté. Les résultats concernant ces items ciblés ont constitué une analyse à part et représentent les principaux résultats de ce travail.

III RESULTATS

3.1 DESCRIPTION GENERALE DES ARTICLES

Les 80 articles se sont répartis comme décrits ci-dessous :

- 26 au total parus dans les revues de médecine nucléaire

Journal of Nuclear Medicine : 3

European journal of Nuclear Medicine and Molecular Imaging : 8

Molecular Imaging : 2

Clinical Nuclear Medicine : 13

- 54 au total parus dans les revues de radiologie

Radiology : 10

Investigative radiology : 2

European Radiology : 18

American journal of Roentgenology : 24

Les articles provenaient majoritairement de centres européens ou américains.

3.2 ITEMS REMARQUABLES (ceux qui n'ont pas été suffisamment informatifs)

Parmi les 29 items de la grille de lecture, 12 ont particulièrement retenu notre attention car ils n'étaient peu ou pas renseignés dans la majorité des articles, ce qui pouvait amener à poser des questions sur la qualité méthodologique. Il s'agit donc d'une piste d'amélioration à étudier. Nous avons choisi de retenir les items pour lesquels le taux d'informativité dans l'ensemble des articles était inférieur à 80%.

Nous allons lister ces items avec leur taux (tableau 2) et les illustrer avec quelques exemples.

NB : quand la discordance entre le junior et sénior n'a pas dépassé 15%, nous avons utilisé le score du sénior sauf dans quelques cas que nous allons développer dans la partie concordance sénior/junior.

Item N.3 : Identifier si les mots clés « agreement » ou « comparison » sont présents dans les mots clés de l'article.

72 articles sur 80 n'ont pas mentionné l'un ou l'autre de ces deux termes parmi leurs mots-clés soit un taux d'absence de 90%.

Comme nous avons précisé plus haut : les bases de données bibliographiques (notamment la Medline via Pubmed) et internet sont devenus les ressources primaires pour la recherche dans une démarche de "Evidence Based Medicine". Pour utiliser efficacement ces bases de données, nous recommandons de bien spécifier (au moins un) les mots clés "agreement" ou "comparison" à la fois dans le titre, le résumé et les mots-clés fournis par le texte. Or même si ces articles traitent de la problématique de la concordance, l'absence de mots clés « agreement » ou « comparison » ne permettra de les identifier lors de la préparation d'une méta-analyse par exemple.

A titre d'exemple, nous avons choisi l'article "Comparison Between ^{99m}Tc-Diphosphonate Imaging and MRI With Late Gadolinium Enhancement in Evaluating Cardiac Involvement in Patients With Transthyretin Familial Amyloid Polyneuropathy" [26].

Ses mots clés officiels sont: ^{99m}Tc-diphosphonate scintigraphy, cardiac amyloidosis, late gadolinium enhancement, MRI, transthyretin familial amyloid polyneuropathy.

Item N.4 : Les objectifs de l'étude portent-ils bien sur la concordance?

35 articles sur 80 qui mentionnaient le terme de concordance dans leur résumé ne portaient pas sur cette problématique de concordance, soit un pourcentage de 44%.

La source la plus fréquente de cette discordance tient au fait qu'ils s'agissaient d'études visant à évaluer la performance diagnostique de la technique (sensibilité, spécificité, valeur prédictive positive et valeur prédictive négative) ou de travaux s'appuyant sur des analyses de simple corrélation.

A titre d'exemple, nous citerons l'article "Dynamic Contrast-Enhanced MR renography for Renal Function evaluation in Ureteropelvic junction obstruction: Feasibility Study" [27].

Il s'agit d'une étude soulevant plusieurs problèmes méthodologiques :

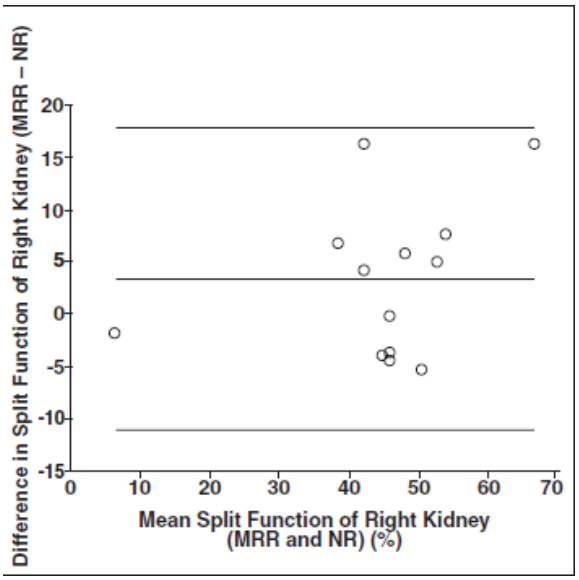
- Elle se présente comme une étude de concordance ayant pour objectif de démontrer que l'IRM peut se substituer à la scintigraphie rénale. Mais l'analyse principale repose sur une étude de corrélation entre les données de l'IRM etles données de la scintigraphie rénale avec un score de Pearson à ($r = 0,87$, $p < 0,01$).

- La taille de l'échantillon est très petite (17 patients) et non justifiée dans le chapitre Matériel et méthodes de l'article.

- La représentation graphique Bland-Altman montre des valeurs dispersées

En effet, les limites de distribution des différences entre les deux méthodes sont très larges alors que le nombre d'observations dans l'échantillon est faible. Certaines des valeurs sont assez éloignées de la différence moyenne (improprement appelé ligne de biais). De plus, la répartition des valeurs expérimentales n'est pas optimale car ne recouvrant pas de façon

homogène le champ des valeurs possibles.



Malgré ces biais importants, les auteurs ont néanmoins conclu à la possibilité de substitution entre les deux méthodes.

Item N 12 : L'expérience des évaluateurs est-elle bien décrite ?

Nous avons choisi de coter l'information donnée en 3 classes :

- bien décrit (avec nombre d'années d'expérience) ;
- peu décrit (mention seulement « expérimenté »)
- non décrit

Dans 19 articles des 80 articles, cet item est peu décrit (soit 24%) et 16 articles ne donnent aucune information « non décrit » (20%).

Quelques exemples d'articles ci-dessous étoffent cette faiblesse de l'information fournie :

- « Prospective evaluation of 68Ga-DOTANOC PET-CT in differentiated thyroid cancer patients with raised thyroglobulin and negative 131I-whole body scan: comparison with 18F-FDG PET-CT »[28]: "*All PET-CT studies were evaluated independently by two experienced Nuclear Medicine physicians."*

- « First imaging results of an intra-individual comparison of 11C-acetate and 18F-fluorocholine PET/CT in patients with prostate cancer at early biochemical first or second relapse after prostatectomy or radiotherapy»[29]: *"PET/CT scans were blindly reviewed by two independent pairs of two experienced nuclear medicine physicians*".

- « Initial clinical results of simultaneous 18F-FDG PET/MRI in comparison to 18F-FDG PET/CT in patients with head and neck cancer »[30]: *"Two reader groups, each composed of one radiologist and one nuclear physician".*

- «Ultrasound Evaluation of gallbladder Dyskinesia: Comparison of Scintigraphy and Dynamic 3D and 4D Ultrasound Techniques» [31]: *"20 healthy volunteers underwent scanning by two sonographers who measured gallbladder volumes with 3D and 4D ultrasound."*

A contrario, l'article suivant donne l'information pertinente sur ce point :

- «Evaluation of Rotator Cuff tendon Tears: Comparison of multidetector CT Arthrography and 1.5-T MR Arthrography»[32]: *"CT and MR arthrographic images were analyzed independently by two fellowship- trained radiologists (F.E.L., and P.O., with 10 and 2 years of experience, respectively, in musculoskeletal radiology"*

Item N 13 : Les évaluateurs sont-ils issus du même centre ?

79 articles sur 80 ne l'ont pas mentionné explicitement si les évaluateurs sont issus ou non du même centre.

Un seul article le précise.

- «The diagnostic value of adding dynamic scintigraphy to standard delayed planar imaging for sentinel node identification in melanoma patients » : [33]. *"images was made as a consensus interpretation by two readers [reading site: Rigs hospitalet (RH)] from the Clinic of Nuclear Medicine, RH, Copenhagen, Denmark."*

Il est vraisemblable que dans la plupart des cas, les évaluateurs impliqués dans une étude soient issus du même centre mais nous ce n'est toujours une certitude. Il est aussi courant que certains centres fassent appel à des collègues d'autres centres nationaux ou étrangers pour des procédures de relecture centralisée. L'origine et les modalités de formation des évaluateurs

sont à prendre en compte dans l'analyse d'un éventuel effet centre (effet cluster d'apprentissage) ou de l'extrapolation à d'autres structures. Aussi, il serait souhaitable que les auteurs améliorent l'information sur l'origine des évaluateurs, comme dans l'article cité plus haut.

Item N 14 : Le recueil des données se fait-il de manière prospective ou rétrospective ?

26 articles sur 80 indiquent qu'il s'agit d'une étude rétrospective (soit un pourcentage de 33%) tandis que 11 articles (soit un pourcentage de 14%) ne précisent pas si les études sont prospectives ou rétrospectives.

Item N 15 : Existe-il une justification du calcul de taille d'échantillon ?

79 études ne donnent aucune justification de la taille de l'échantillon inclus dans l'étude pour évaluer la concordance entre les techniques à comparer.

Le seul article qui renseigne cet item est :

- «Patent Foramen Ovale: Detection with Nongated Multidetector CT» [34].

En effet, les auteurs ont réalisé un calcul préalable aboutissant à un nombre de sujet à inclure (n = 107).
Même si cela est rare dans les études diagnostiques, on peut citer un autre exemple publié dans le « JAMA » en 2004 et se féliciter sa qualité méthodologique :

- « Comparaison IRM - Scanner dans la détection des hémorragies cérébrales aiguës » [46]. Ce travail cherchait à évaluer la fiabilité de l'IRM en écho de gradient versus scanner sans injection de produit de contraste dans la détection de l'hémorragie cérébrale aigue. Ainsi les auteurs ont fourni des hypothèses pour le calcul de l'effectif : *" Pour le calcul initial du nombre de sujets nécessaires à l'étude, nous avons supposé que le scanner était un examen fiable à 100 % pour détecter une hémorragie et notre objectif était de démontrer que l'IRM était également fiable à 100 %. Avec ce plan expérimental visant à démontrer une relation de non-infériorité, il fallait, pour limiter à moins de 5 % la différence entre l'IRM et le scanner dans l'intervalle de confiance (IC) à 95 %, une concordance de l'IRM et du scanner pour 55 hémorragies*. "

Item N 18 : le délai entre l'administration du radiopharmaceutique et l'acquisition des images est-il bien précisé (pour les examens de médecine nucléaire) ?

15 articles sur 41 ne précisent pas ce délai (soit un pourcentage de 37%) tandis que les 26 autres articles le précisent.

NB : 39 articles ne sont pas éligibles pour cette analyse puisqu'il s'agit des études évaluant les techniques de radiologie pures sans faire intervenir les examens de médecine nucléaire.

Un exemple où cet item est absent :

- « Utility of Salivagram in Pulmonary aspiration in Pediatric Patients: Comparison of Salivagram and Chest Radiography»: [35]. *"The radiopharmaceutical was administered orally as a small drop (approximately 100 μL) placed with a syringe either on the base of the patient's tongue or in a sublingual location while the patient was lying in a supine position on the imaging table. The patient was allowed to swallow naturally and posterior planar imaging of the mouth, chest, and upper abdomen was obtained. Continuous dynamic 30-second images were recorded for a total of 60 minutes."*

Un exemple où cet item est présent:

- «First imaging results of an intra-individual comparison of 11C-acetate and 18F-fluorocholine PET/CT in patients with prostate cancer at early biochemical first or second relapse after prostatectomy or radiotherapy»: [29]. *"A first whole-body PET study was performed 5 min after tracer injection".*

A signaler pour cet item qu'il ne faut pas se limiter à la présence de cette seule information car il s'agit du délai théorique entre l'administration du radiopharmaceutique et le début des acquisitions. Il faut aussi obtenir le délai réel moyen observé lors de la réalisation du travail.

Item N 20 : L'interprétation est-elle en insu et bien décrite?

Nous avons choisi de coter en 3 classes :

- Insu total (absence de connaissance du contexte clinico-biologique et de résultat d'une autre modalité d'imagerie) ;

- Insu partiel (connaissance d'au moins l'un des éléments cité ci-dessus)

- non décrit (mention seulement «d'insu» sans autre précision)

Vingt sept articles sur 80 s'inscrivent dans la classe non décrit (soit 34%).

Un exemple où cet item est non décrit:

- «Proton MRI in the evaluation of pulmonary sarcoidosis: Comparison to chest CT»: [36]. *"Two cardiopulmonary radiologists, with 15–20 years of radiology experience, then independently reviewed and scored the randomized MRI cases."*

En effet, le terme « independently» signifie qu'ils analysent séparément les images sans établir par la suite un consensus entre eux. Par ailleurs, il n'y a aucune notion de l'insu.

Item N.22 existe-t-il une comparaison par rapport au gold standard ?

35 articles sur 80 (soit 44%) n'ont pas comparé le résultat par rapport au gold standard. Dans la plupart des cas, les auteurs se sont contentés d'un suivi simple du patient ou n'ont pas évoqué une comparaison par rapport au gold standard.

On peut citer des exemples :

- «Comparison of ^{99m}Tc-MIBI SPECT/^{18}F-FDG PET Imaging and Cardiac Magnetic Resonance Imaging in Patients With idiopathic Dilated Cardiomyopathy» [37]. Cet article n'a pas évoqué dans la partie "Materiel and methods" la notion de gold standard

- « ^{68}Ga-DOTA-NOC PET/CT in comparison with CT for the detection of bone metastasis in patients with neuroendocrine tumours» [38]. Dans cet article, les auteurs se contentent d'un suivi clinique simple sans évoquer la possibilité d'une biopsie.

Item N.23 Existe-t-il une description des tests statistiques utilisés ?

Tous les articles contiennent un chapitre où est présentée la méthodologie statistique prévue. Nous nous sommes alors concentrés sur la qualité de la description.

En effet, 37 articles sur 80 (soit 46%) n'ont pas décrit le type de test statistique utilisé. Comme évoqué dans l'introduction, le choix du test dépend du contexte et de la question étudiée et devrait comporter le plus de détails possibles.

Par exemples, les auteurs se contentent de mentionner la statistique kappa de Cohen sans plus de précision [39-43]. Or il en existe plusieurs variantes dont notamment le kappa pondéré [44].

Item N.28 Existe-t-il une ou des déviation(s) au protocole?

Cinq articles sur 80 (soit 6%) ont signalé une déviation au protocole tandis que 74 (soit 93%) articles n'évoquent pas ce point. Bien que nous pensons que la plupart des études ont respecté le protocole initial, il serait préférable de le mentionner explicitement dans la partie « discussion ».

Voici quelques exemples des études qui l'ont signalé :

- «Noninvasive evaluation of active lower gastrointestinal bleeding: comparison between contrast-enhanced MDCT and 99mTc-labeled RBC scintigraphy.» [45]. En effet, parmi les 60 patients inclus, 41 patients seulement ont effectué la scintigraphieaux hématies marquées. Parmi les 19 patients qui n'ont pas suivi le protocole, les auteurs ont décrit la cause de la déviation au protocole dans 5 cas (hémorragie menaçante ayant nécessité le recours de radiologie interventionnelle.

- «Femoral and Tibial Torsion measurements With 3D Models based on Low-Dose Biplanar radiographs in Comparison With standard CT Measurements» [46].

En effet dans cet article, contrairement à ce qui était initialement annoncé, les auteurs n'ont pas correctement effectué l'acquisition d'image chez un patient (problème de positionnement).

Item N.29 L'interprétation est-elle pertinente cliniquement?

Aucune étude, dans sa partie discussion, n'a discuté la pertinence clinique de l'interprétation. Tous les auteurs se sont appuyés sur des données communément admises (cf la partie sur la notion de concordance).

Rappelons qu'il n'y a pas de chiffre « significatif» ou non pour la concordance (coefficient de Kappa ou ICC par exemples) puisqu'il s'agit d'une mesure de l'accord entre observateurs, qui doit être interprétée en fonction du contexte.

Il serait souhaitable d'intégrer cette réflexion basée sur les résultats de l'étude pour appliquer dans le contexte clinique pour mieux sélectionner un éventuel examen de substitution.

3.3 AUTRES ITEMS (ceux qui ont été suffisamment informatifs)

Les autres items (18) de la liste ne sont pas aussi importants pour ce travail car la majorité d'articles les a pris en compte : nous allons lister ces items avec leur pourcentage de complétion (tableau 2) et les illustrer avec quelques contre-exemples.

NB : quand la discordance entre le junior et sénior n'a pas dépassé 15%, nous avons utilisé le score du sénior sauf dans quelques cas que nous allons développer dans la partie concordance sénior/junior.

Item N.1 : Identifier si les mots clés « agreement » ou « comparison » sont présents dans le titre

75 articles sur 80 (soit 94%) respectent ce point.

Contre-exemple: « Bladder Cancer: Diagnosis with Diffusion-weighted MR Imaging in Patients with Gross Hematuria» [47]

Item N.2 : Identifier si les mots clés « agreement » ou « comparison » sont présents dans le résumé

80 articles sur 80 (soit 100%) respectent ce point.

Item N.5 : Les méthodes de mesure sont-elles explicitement et précisément nommées (pas seulement en abréviation) et décrites?

72 articles sur 80 (soit 91%) respectent ce point.

Contre-exemple: « Dynamic Contrast-Enhanced MR renography for Renal Function evaluation in Ureteropelvic junction obstruction: Feasibility Study. AJR Am J Roentgenol. » [27]

Item N.6 : L'affection est-elle bien définie?

73 articles sur 80 (soit 92%) respectent ce point.

Contre-exemple: « A comparison between NASCET and ECST methods in the study of carotids: Evaluation using Multi-Detector-Row CT angiography » [48]

Item N.7 : Pour la population étudiée: les méthodes d'inclusion sont-elles bien décrites?

73 articles sur 80 (soit 89%) respectent ce point.

Contre-exemple: « Ultrasound evaluation of gallbladder dyskinesia: comparison of scintigraphy and dynamic 3D and 4D ultrasound techniques. » [31]

En effet, les critères de non-inclusion n'ont pas été mentionnés.

Item N.8 : Pour l'échantillon de la population étudiée : le nombre de sujets dans l'échantillon étudié est-il défini ?

80 articles sur 80 (soit 100%) respectent ce point.

Item N.9: Pour l'échantillon de la population étudiée: avons-nous l'information précise sur la distribution de l'âge (moyenne et médiane) dans la population étudiée ?

76 articles sur 80 (soit 95%) respectent ce point.

Contre-exemple: « Utility of salivagram in pulmonary aspiration in pediatric patients: comparison of salivagram and chest radiography. AJR Am J Roentgenol» [35]

En effet, les auteurs se sont contentés de parler de sujets qui ont « moins de 21 ans ».

Item N.10 : Avons-nous l'information précise sur le sexe ratio?

77 articles sur 80 (soit 96%) respectent ce point.

Contre-exemple: « Utility of salivagram in pulmonary aspiration in pediatric patients: comparison of salivagram and chest radiography. AJR Am J Roentgenol» [35]

Item N 11 : Le nombre des évaluateurs est-il bien précisé ?

78 articles sur 80 (soit 98%) respectent ce point.

Contre-exemple: «A prospective study comparing whole-body FDG PET/CT to combined planar bone scan with 67Ga SPECT/CT in the Diagnosis of Spondylodiskitis» [49]

Item N 16 : Existe-il une description précise du déroulement de l'examen ?

80 articles sur 80 (soit 100%) respectent ce point.

Item N 17 : L'activité administrée est-elle bien précisée (pour les examens de médecine nucléaire) ?

Tous les articles soit 41 articles sur 41 (soit 100%) respectent ce point.

NB : 39 articles ne sont pas éligibles pour cette analyse car il s'agit des études évaluant les techniques de radiologie pures ne faisant pas intervenir les examens de médecine nucléaire.

Item N 19 : Le protocole d'acquisition des images est-il bien précisé ?

78 articles sur 80 (soit 98%) respectent ce point.

Contre-exemple: «Whole-body MRI with diffusion-weighted sequence for staging of patients with suspected ovarian cancer: a clinical feasibility study in comparison to CT and FDG-PET/CT» [50].

Item N 21 : Existe-il une description de la procédure de définition des classes de résultats (binaire, ordinal, classe, variable quantitative ect..) ?

80 articles sur 80 (soit 100%) respectent ce point.

Item N.25 Est-ce que le nombre de sujets analysés est identique à celui du nombre de sujets inclus dans l'étude?

75 articles sur 80 (soit 94%) respectent ce point.

Item N.26 Existe-t-il un tableau récapitulant les motifs de sortie de l'essai ?

Les 5 articles qui n'ont pas respecté le point N.25 respectent ce point.

Prenons l'exemple de l'article « Evaluation of angiographic computed tomography in the follow-up after endovascular treatment of cerebral aneurysms--a comparative study with DSA and TOF-MRA» [51]

En effet dans cet article, un patient a été exclu de l'analyse car un clip chirurgical intracérébral avait été posé entrainant une extinction de signal en Angio-IRM avec TOF, ce qui n'avait pas été prévu au début de l'étude.

Item N.27 Existe-t-il une description des caractéristiques démographiques des sujets inclus ?

67 articles sur 80 (soit 83%) respectent ce point.

Il est préférable de mettre un tableau résumant les caractéristiques démographiques (qui est un critère essentiel pour comparer des travaux sur la même technique mais avec des populations différentes (biais de spectre). Cette information est aussi nécessaire pour savoir si les résultats d'un travail sont extrapolables ou non à la population globale des patients.

3.4 CONCORDANCE SENIOR/JUNIOR

Il existait une bonne concordance senior/junior puisque la plupart des items ont eu une concordance supérieure à 80% (tableau 3).

Nous allons lister les ceux qui ont eu une concordance moins bonne et tenter de proposer des explications :

- Item N.4 : Les objectifs de l'étude portent-ils bien sur la concordance?

Il existe une concordance pour 57 articles sur 80 (soit 73%).

Il s'agit d'une lecture différente de l'item. La principale discordance résidait dans le fait que le junior cotait l'item présent lorsque la concordance fait partie du résultat de l'étude (notamment dans certaines études où le résultat du critère de jugement principal n'est pas explicitement mentionné). Tandis que le sénior était plus tolérant lorsque les performances diagnostiques faisaient partie du résultat principal de l'étude.

- Item N.7 : Pour la population étudiée: les méthodes d'inclusion sont-elles bien décrites?

Il existe une concordance pour 61 articles sur 80 (soit 77%).

Il s'agit aussi d'un item assez subjectif : le junior a côté item présent lorsque les critères d'inclusion et de non-inclusion étaient présents, sans trop aborder l'analyse qualitative qui est de savoir si ces critères étaient pertinents pour la question étudiée.

- Item N 17 : L'activité administrée est-elle bien précisée (pour les examens de médecine nucléaire et les examens radiologiques) ?

Il existe une concordance de 75 sur 80 (soit 94%) alors la concordance initiale était pour 55 articles sur 80 (soit 69%).

En effet, le sénior a côté l'item présent lorsqu'il s'agissait aussi d'un examen de radiologique avec injection de produit de contraste. Le junior voulait initialement que la quantité injectée de produit de contraste ne faisait pas partie de l'item. Finalement après la discussion, compte tenu de l'importance de cette mention notamment en ce qui concerne les effets indésirables liés au produit de contraste, nous avons retenu cet item pour les examens de médecine nucléaire et les examens radiologiques.

Concernant les items N.28 (Existe-t-il une déviation au protocole?) et N.29 (Existe-t-il une pertinence clinique de l'interprétation ?), la concordance initiale a été très mauvaise (moins de 20%). Après une réunion dans laquelle le junior a exposé ses explications concernant le codage, la concordance s'est beaucoup améliorée.

En effet concernant l'item N.28 : le sénior considérait que l'absence d'évocation dans la discussion signifiait qu'il n'existe pas de déviation de protocole. Cela est probablement vrai dans la plupart des articles. Mais le junior a insisté au nom d'une méthodologie rigoureuse que les auteurs devraient systématiquement mentionner l'existence ou non de déviation au protocole.

Concernant l'item N.29, le concept n'avait pas été suffisamment précisé avant le début de l'analyse. Le junior avait privilégié le fait qu'il ne suffisait pas de donner une interprétation seulement en fonction du résultat numérique obtenu mais qu'il était souhaitable de commenter le résultat en intégrant le contexte. Au contraire, le sénior avait favorisé la pertinence globale de l'article. Lors de la réunion après analyse de 60 articles, la méthode du junior a finalement était retenu.

Prenons plusieurs exemples :

Dans l'article « A comparison between NASCET and ECST methods in the study of carotids evaluation using Multi-Detector-Row CT angiography» [48]

Le senior a considéré que ce point a été respecté puisque les auteurs ont conclu à une bonne concordance (good agreement) entre les deux techniques en se basant sur la valeur de kappa égale à 0,825.

De même, dans l'article « PET/CT in Lymphoma: Prospective Study of Enhanced Full-Dose PET/CT Versus Unenhanced Low-Dose PET/CT» [52] :

Le senior a également considéré que ce point avait été respecté puisque les auteurs ont conclu à une concordance quasi parfaite « almost perfect agreement » en se basant sur la valeur de kappa égale à 0,92.

Ces deux articles sont de bonne qualité sur le plan méthodologique et apporte des preuves robustes sur la possibilité de substitution de techniques. Cependant, le junior a estimé qu'ils n'ont pas respecté l'item 29 car les auteurs n'ont pas intégré les données du contexte clinique aux résultats afin de mieux évaluer l'éventuelle substitution par un autre examen.

Dans l'article « Whole-Body 18F-Fluorocholine (FCH) PET/CT and MRI of the Spine for Monitoring Patients With Castration-Resistant prostate Cancer Metastatic to Bone »[53], le senior a considéré que ce point n'avait pas été respecté en raison de la taille de l'échantillon qui était très faible (10 cas). Il s'agit plutôt d'un problème global de méthodologie de l'étude et non d'un problème spécifique de l'item.

4. DISCUSSION

4.1 REPRESENTATIVITE ET GENERALISABILITE DES RESULTATS RAPPORTES DANS CE TRAVAIL

Pour garantir le caractère exhaustif de cette étude, nous nous sommes posés la question du biais de sélection, notamment concernant la sélection des revues de spécialité. Finalement, la recherche s'est effectuée sur les revues de spécialité de l'imagerie médicale. Il serait fort utile d'analyser un échantillon d'études d'imagerie publiées dans des revues de spécialité médicale ou chirurgicale (Par exemple une étude de scintigraphie rénale publiée dans un Journal de Pédiatrie). Il est possible que le respect des items proposés soit beaucoup moins fiable dans ces revues où les reviewers sont moins sensibilisés aux spécificités de l'imagerie médicale.

Ainsi un travail de vérification a été effectué au courant septembre 2014 avec l'aide de M. CHEKIB Vincent. Nous avons sélectionné un échantillon de revues ayant des impacts factors les plus élevés dans quelques disciplines.

Ces revues étaient récentes couvrant une année (du 01.01.2010 au 31.12.2010).

Au final, la plupart des revues ont moins de 5% d'articles traitant spécifiquement l'imagerie médicale. (Fig 3)

Nous pouvons donc affirmer la bonne représentativité de cette étude.

4.2 HYPOTHESES SUR LES ITEMS REMARQUABLES

Nous essayons de comprendre pourquoi il existait quelques items où le taux était très bas.

Ces items concernent :

Item N.13 : Les évaluateurs sont-ils issus du même centre ? (1%)

Item N.15 : Existe-il un calcul de taille d'échantillon ? (1%)

Item N.28 Existe-t-il une ou des déviation(s) au protocole? (7%)

Item N.29 Existe-t-il une pertinence clinique de l'interprétation ? (0%)

Nous avons émis certaines hypothèses :

- Question d'habitudes /culture (concernant les item N.13 et N.28) : les auteurs n'ont pas l'habitude de préciser si les évaluateurs sont-ils issus du même centre ou s'il existe une déviation au protocole. En effet, il semblerait dans la plupart d'études, ces deux points sont relativement bien respectés et ils sont tellement évidents que les auteurs n'ont pas pensé à préciser. Or pour avoir une méthodologie rigoureuse, il est nécessaire de disposer d'une grille standardisée dans la rédaction d'article. Nous recommandons vivement ces deux points soient intégrés et bien respectés.

- Problème de bonne pratique (concernant l'item N.15) : il semblerait que la plupart d'études diagnostiques concernant les imageries médicales ont une méthodologie moins rigoureuse que les études concernant les essais thérapeutiques (où la grille de rédaction est beaucoup mieux codifiée). L'absence quasi-systématique de cet item n'est pas un hasard mais plutôt un point à améliorer.

- Problème de prise de conscience : L'absence systématique de l'item 29 reflète particulièrement le problème de prise de conscience. Les auteurs ont tous commenté leur résultat de concordance à l'aide des seuils qui sont communément admis dans la littérature. Or il s'agit des seuils qui ont été définis de façon arbitraire (comme c'est le cas de la valeur de « p »). Cependant, il serait souhaitable de moduler ces seuils en fonction du contexte. En effet, comme nous l'avons expliqué précédemment : s'il s'agit d'un examen qui a une valeur décisive sur le pronostic vital ou fonctionnel du patient, l'exigence du niveau de concordance sera extrêmement important, quasi parfait. Autrement dit, il s'agit d'être strict vis-à-vis du choix de l'examen de substitution dans les situations à risque. Nous pensons qu'il est souhaitable que les auteurs abordent cet aspect dans la discussion.

Concernant les autres items remarquables, nous s'orientons plutôt vers une hétérogénéité de qualité méthodologique pour laquelle il existe un vrai effort à fournir.

Le reste des items que la plupart des articles respectent bien, ne nécessite pas un effort d'amélioration particulier.

4.3 HYPOTHESE SUR LES DISCORDANCES (SENIOR/JUNIOR)

La concordance générale est bonne (plus part de cas supérieur à 80%). Cela témoigne qu'en général avec une lecture sérieuse et méthodique, deux personnes avec un niveau d'expérience assez différent peuvent avoir un résultat similaire. Il serait souhaitable qu'on consacre un minimum temps de lecture (en général 1h par article au début de l'analyse et vers 30 minutes en moyenne vers la fin d'analyse, en rapport avec l'effet d'apprentissage).

Les principales discordances observées sont liées :

- d'une part à la compréhension : en effet certains articles n'étaient pas clairs dans la rédaction, générant ainsi la difficulté de compréhension et favorisant une interprétation subjective.

- d'autre à l'expérience et au niveau de formation : dans quelques cas, le junior n'est montré moins à l'aise pour interpréter les parties « résultat » et « discussion » en raison d'une expérience nettement moindre dans l'exploitation des données de la littérature médicale.

4.4 PORTEE DE L'ETUDE

Comme cela a été rappelé dans l'introduction, ce travail s'inscrit dans le contexte d'une prise de conscience par la communauté médicale et scientifique des enjeux du taux encore élevé de non reproductibilité dans les études cliniques. Nous avons constaté une nette volonté d'amélioration notamment de la part des organismes de tutelle. En effet, dans le site dédié aux appels d'offres de PHRC de Grand Ouest, il existe des recommandations de rédaction de protocole destinées aux investigateurs parmi lesquels un document intitulé "Plans types à décliner pour essai portant sur un dispositif médical" présente quelques similarités avec ce travail.

Nous pouvons citer quelques exemples:

- la section "4.2.2 Déroulement de l'étude" présente une recommandation ". Ce paragraphe doit décrire les méthodes de mesure des paramètres étudiés le plus précisément possible :

« Méthode de mesure, appareil utilisé, produit utilisé, fournisseur, lieu de mesure, responsable des mesures… » qui correspond à l'esprit des items N.16 et N.19

- la section "4.4.2 Méthodes de mise en insu" présente une recommandation "Méthodes mises en place pour l'aveugle du patient, de l'investigateur (care provider) et de l'évaluateur (outcome assessor)" qui s'apparente à l'item N. 20

- la section "5.5.4 Modalités de suivi de ces personnes" présente une recommandation "Définir le suivi qui doit être mis en place pour un sujet sorti d'étude, ainsi que les données devant être recueillies" qui correspondent aux items 25 et 26.

- la section "6.1.1.4 Résumé de la formation et de l'expérience nécessaires à l'utilisation du dispositif médical ou dispositif médical de diagnostic in vitro" s'apparente à l'item N.13

- la section "9.1 Description des méthodes statistiques prévues, y compris du calendrier des analyses intermédiaires prévues : Spécifier les méthodes statistiques qui seront utilisées" s'apparente aux items N.23 et 24.

- la section "9.2 Nombre prévu de personnes à inclure dans la recherche, et nombre prévu de personnes dans chaque lieu de recherche avec sa justification statistique" s'apparente à l'item N.15

- la section "9.5 Méthode de prise en compte des données manquantes, inutilisées ou non valides" s'apparente aux items N.25 et 26.

Il est aussi à souligner le fait que plus d'un quart d'articles sélectionnés (24 sur 80) est issu de la revue « American journal of Roentgenoloy » dans laquelle nous avons constaté qu'il existe une proportion de défaut méthodologique plus élevée par rapport aux autres revues. Paradoxalement, cette prépondérance en nombre d'articles publiés contraste avec leur qualité méthodologique relativement médiocre, ce qui est problématique puisqu'il s'agit d'un aspect important pour juger la validité de l'étude. Nous espérons que les relecteurs de certaines revues soient plus exigeants quant à la sélection d'articles.

5. CONCLUSION :

Le problème de non reproductibilité est réel et nécessite un effort collectif. Les études d'imagerie diagnostique ne font pas exception. Cette étude, bien que modeste, essaie de participer à l'amélioration de Bonnes Pratiques médicale en proposant quelques réflexions et certains axes d'amélioration.

Après analyse de 29 items, 12 en particulier sont potentiellement améliorables. Il s'agit notamment des items concernant la justification de la taille de l'échantillon, le centre dont est issu l'évaluateur, la précision quant à l'existence d'une déviation au protocole et la pertinence du résultat de concordance intégrée dans le contexte clinique.

La concordance senior-junior est bonne, suggérant que les expériences dans la pratique médicale et l'habitude à lire les articles n'ont pas d'influence décisive dans la compréhension d'un article sous réserve que le temps de lecture soit suffisant. Néanmoins nous avons remarqué qu'un consensus bien établi avant le début d'analyse et une discussion permanente pendant l'analyse faciliteraient le travail augmentant ainsi la concordance.

Annexe 1

Figure 1 : Diagramme de flux concernant la sélection des articles

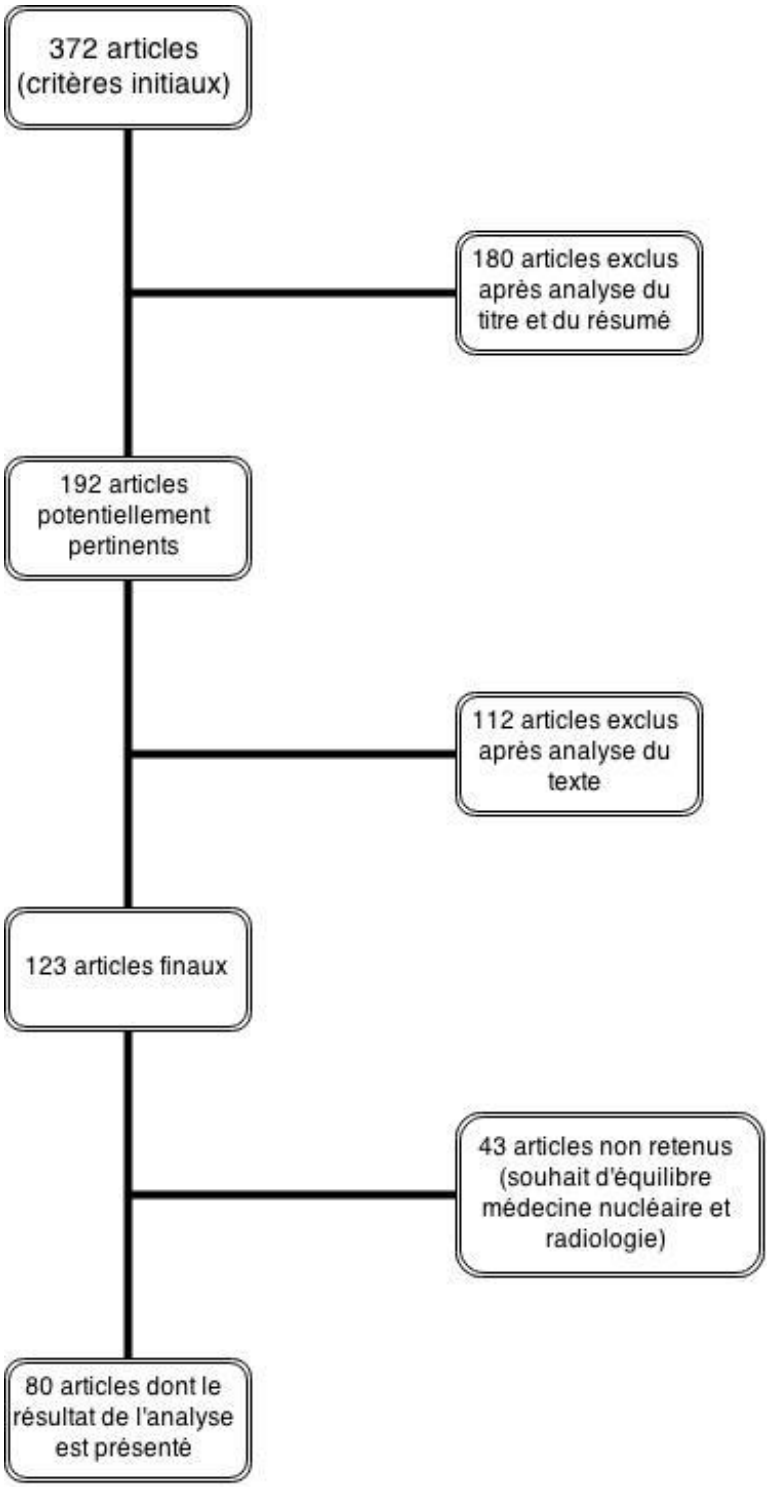


Fig 2 : La page montrant la sélection des 372 articles répondant aux critères de recherche initiaux.

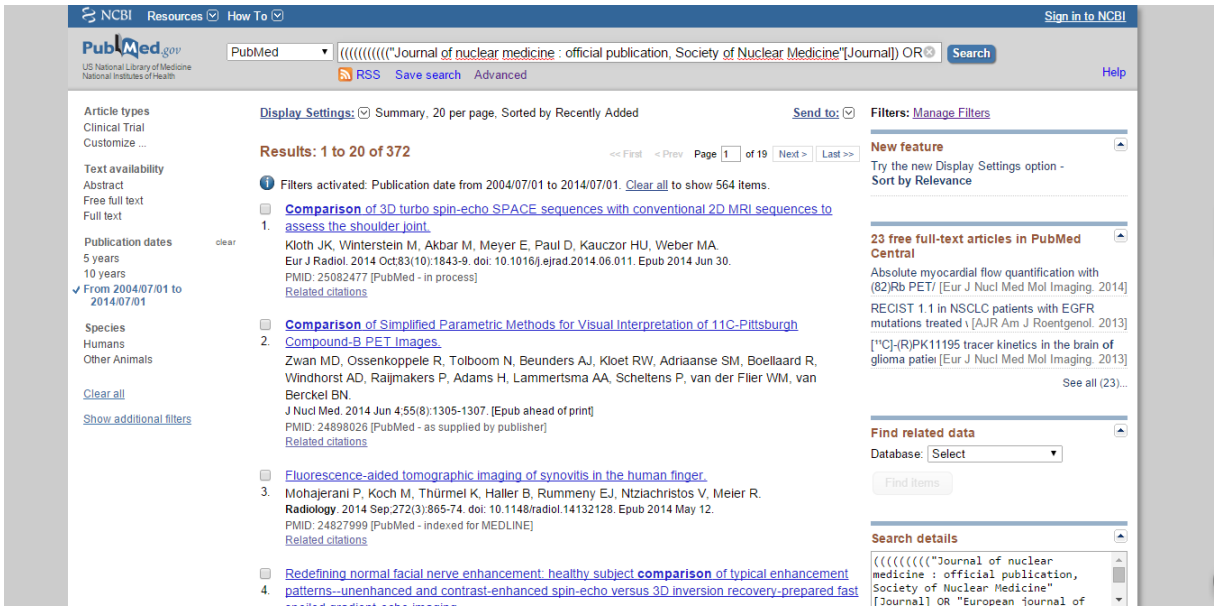


Fig 3 : Pourcentage des articles sur l'imagerie médicale dans quelques revues phares (ne traitant pas d'imagerie médicale) en 2010

	A	B	C	D	E	F	G
1		Nbre articles en 2010	Nbre articles en 2010 sur l'imagerie médicale	ıédicale en 2010	Nbre articles en 2011	Nbre articles en 2011 sur l'imagerie médicale	Pourcentage des articles sur l'imagerie médicale en 2011
2	Lancet	1539	40	3	1543	71	5
3	British medical journal	2096	62	3	2604	77	3
4	JAMA	1042	37	4	1011	46	5
5	The New England journal of medicine	1403	121	9	1417	92	6
6	Circulation	906	190	21	966	160	17
7	Cell Metabolism	155	4	3	185	8	4
8	Blood	1951	63	3	2178	83	4
9	PLoS medicine	196	3	2	204	4	2
10	Endocrine reviews	39	0	0	30	0	0
11							
12							
13	Equation de recherche pour le nombre d'articles publiés dans une revue donnée durant une année :						
14							
15	**lancet[Journal] AND ("2010/01/01"[PDAT] : "2010/12/31"[PDAT])**						
16							
17	Equation de recherche pour le nombre d'articles sur l'imagerie médicale publiés dans une revue donnée durant une année :						
18							
19	**diagnostic imaging[MeSH Terms] AND "lancet"[Journal] AND ("2010/01/01"[PDAT] : "2010/12/31"[PDAT])**						
20							
21							
22							
23							

Annexe 2

TABLEAU 1: Les différents items

Item	Parties
Item N.1 : Identifier si les mots clés « agreement » ou « comparison» sont présents dans le titre	**TITRE ET RESUME**
Item N.2 : Identifier si les mots clés « agreement » ou « comparison » sont présents dans le résumé	
Item N.3 : Identifier si les mots clés « agreement » ou « comparison » sont présents dans les mots clés de l'article	
Item N.4 : Les objectifs de l'étude portent –ils bien sur la concordance?	**INTRODUCTION**
Item N.5 : Les méthodes de mesure sont-elles explicitement et précisément nommées (pas seulement en abréviation) et décrites?	
Item N.6 : L'affection est-elle bien définie?	
Item N.7 : Pour la population étudiée: les méthodes d'inclusion sont-elles bien décrites?	**MATERIELS ET METHODES**
Item N.8 : Pour la population étudiée : le nombre de sujets de la population étudiée est-il bien défini ?	
Item N.9: Pour l'échantillon de la population étudiée: avons-nous l'information précise sur la distribution de l'âge (moyenne et médiane) dans la population étudiée ?	
Item N.10 : Avons-nous l'information précise sur le sexe ratio?	
Item N. 11 : Le nombre des évaluateurs est-il bien précisé ?	
Item N. 12 : L'expérience des évaluateurs est-elle bien décrite ?	

Item N. 13 : Les évaluateurs sont-ils issus du même centre ?	
Item N. 14 : Le recueil des données se fait-il de manière prospective ou rétrospective ?	
Item N. 15 : Existe-il une justification du calcul de taille d'échantillon ?	
Item N. 16 : Existe-il une description précise du déroulement de l'examen ?	
Item N. 17 : L'activité administrée est-elle bien précisée ?	
Item N. 18 : Le délai entre l'administration du radio-pharmaceutique et l'acquisition des images est-il bien précisé?	
Item N. 19 : Le protocole d'acquisition des images est-il bien précisé ?	
Item N. 20 : L'interprétation est-elle en insu et bien décrite?	
Item N. 21 : Existe-il une description de la procédure de définition des classes de résultats (binaire, ordinal, classe, variable quantitative ect..) ?	
Item N. 22 Existe-t-il une comparaison par rapport au gold standard ?	
Item N. 23 Existe-t-il une description des tests statistiques utilisés ?	
Item N. 24 Ces tests statistiques sont-ils pertinents?	
Item N. 25 Est-ce que le nombre de sujets analysés est identique par rapport au nombre de sujets inclus?	**RESULTATS ET DISCUSSION**
Item N. 26 Existe-t-il un tableau récapitulant les motifs de sortie de l'essai ?	
Item N. 27 Existe-t-il une description des caractéristiques démographiques des sujets inclus ?	
N. 28 Existe-t-il une ou des déviation(s) au protocole?	

N. 29 Existe-t-il une pertinence clinique de l'interprétation ?

TABLEAU 2 : Taux de présence des items (soulignés : items remarquables)

Item	Taux de présence
Item N.1 : Identifier si les mots clés « agreement » ou « comparison» sont présents dans le titre	**94%**
Item N.2 : Identifier si les mots clés « agreement » ou « comparison » sont présents dans le résumé	**100%**
Item N.3 : Identifier si les mots clés « agreement » ou « comparison » sont présents dans les mots clés de l'article	**10%**
Item N.4 : Les objectifs de l'étude portent-ils bien sur la concordance?	**56%**
Item N.5 : Les méthodes de mesure sont-elles explicitement et précisément nommées (pas seulement en abréviation) et décrites?	**91%**
Item N.6 : L'affection est-elle bien définie?	**92%**
Item N.7 : Pour la population étudiée: les méthodes d'inclusion sont-elles bien décrites?	**89%**
Item N.8 : Pour la population étudiée : le nombre de sujets de la population étudiée est-il bien défini ?	**100%**
Item N.9: Pour la population étudiée: avons-nous l'information précise sur la distribution de l'âge (moyenne et médiane) dans la population étudiée ?	**95%**
Item N. 10 : Avons-nous l'information précise sur le sexe ratio?	**96%**
Item N. 11 : Le nombre des évaluateurs est-il bien précisé ?	**98%**

Item N. 12 : L'expérience des évaluateurs est-elle bien décrite ?	**56%**
Item N. 13 : Les évaluateurs sont-ils issus du même centre ?	**1%**
Item N. 14 : Le recueil des données se fait-il de manière prospective ou rétrospective ?	**53% (prospective)**
Item N. 15 : Existe-il une justification du calcul de taille d'échantillon ?	**1%**
Item N. 16 : Existe-il une description précise du déroulement de l'examen ?	**100%**
Item N. 17 : L'activité administrée est-elle bien précisée ?	**100%**
Item N. 18 : Le délai entre l'administration du radio-pharmaceutique et l'acquisition des images est-il bien précisé?	**63%**
Item N. 19 : Le protocole d'acquisition des images est-il bien précisé ?	**98%**
Item N. 20 : L'interprétation est-elle en insu et bien décrite?	**66%**
Item N 21 : Existe-il une description de la procédure de définition des classes de résultats (binaire, ordinal, classe, variable quantitative ect..) ?	**100%**
Item N.22 Existe-t-il une comparaison par rapport au gold standard ?	**56%**
Item N.23 Existe-t-il une description des tests statistiques utilisés ?	**54%**
Item N. 24 Ces tests statistiques sont-ils pertinents?	**100%**
Item N.25 Est-ce que le nombre de sujets analysés est identique par rapport au nombre de sujets inclus?	**94%**
Item N.26 Existe-t-il un tableau récapitulant les motifs de sortie de l'essai ?	**100%**
Item N.27 Existe-t-il une description des caractéristiques démographiques des sujets inclus ?	**83%**

Item N.28 Existe-t-il une ou des déviation(s) au protocole?	**7%**
Item N.29 Existe-t-il une pertinence clinique de l'interprétation ?	**0%**

TABLEAU 3 : Concordance Sénior/Junior

Item	Taux de concordance
Item N.1 : Identifier si les mots clés « agreement » ou « comparison» sont dans le titre	**94%**
Item N.2 : Identifier si les mots clés « agreement » ou « comparison » sont dans le résumé	**100%**
Item N.3 : Identifier si les mots clés « agreement » ou « comparison » sont présents dans les mots clés de l'article	**100%**
Item N.4 : Les objectifs de l'étude portent-ils bien sur la concordance?	**73%**
Item N.5 : Les méthodes de mesure sont-elles explicitement et précisément nommées (pas seulement en abréviation) et décrites?	**91%**
Item N.6 : L'affection est-elle bien définie?	**90%**
Item N.7 : Pour la population étudiée: les méthodes d'inclusion sont-elles bien décrites?	**77%**
Item N.8 : Pour la population étudiée : le nombre de sujets de la population étudiée est-il bien défini ?	**98%**
Item N.9: Pour la population étudiée: avons-nous l'information précise sur la distribution de l'âge (moyenne et médiane) dans la population étudiée ?	**95%**
Item N. 10 : Avons-nous l'information précise sur le sexe ratio?	**96%**
Item N. 11 : Le nombre des évaluateurs est-il bien précisé ?	**98%**

Item N. 12 : L'expérience des évaluateurs est-elle bien décrite ?	**88%**
Item N. 13 : Les évaluateurs sont-ils issus du même centre ?	**98%**
Item N. 14 : Le recueil des données se fait-il de manière prospective ou rétrospective ?	**88%**
Item N. 15 : Existe-il une justification du calcul de taille d'échantillon ?	**100%**
Item N. 16 : Existe-il une description précise du déroulement de l'examen ?	**95%**
Item N. 17 : L'activité administrée est-elle bien précisée ?	**69%**
Item N. 18 : Le délai entre l'administration du radio-pharmaceutique et l'acquisition des images est-il bien précisé?	**80%**
Item N. 19 : Le protocole d'acquisition des images est-il bien précisé ?	**89%**
Item N 20 : L'interprétation est-elle en insu et bien décrite?	**88%***
Item N. 21 : Existe-il une description de la procédure de définition des classes de résultats (binaire, ordinal, classe, variable quantitative ect..) ?	**89%**
Item N.22 Existe-t-il une comparaison par rapport au gold standard ?	**88%**
Item N.23 Existe-t-il une description des tests statistiques utilisés ?	**98%**
Item N. 24 Ces tests statistiques sont-ils pertinents?	**99%**
Item N.25 Est-ce que le nombre de sujets analysés est identique par rapport au nombre de sujets inclus?	**83%**
Item N.26 Existe-t-il un tableau récapitulant les motifs de sortie de l'essai ?	**95%**
Item N.27 Existe-t-il une description des caractéristiques démographiques des sujets inclus ?	**89%**

Item N.28 Existe-t-il une ou des déviation(s) au protocole?	**95%***
Item N.29 Existe-t-il une pertinence clinique de l'interprétation ?	**100%***

Annexe 3

LEXIQUE (référence : Guidance for Industry "VALIDATION OF ANALYTICAL PROCEDURES: DEFINITION AND TERMINOLOGY")

ACCURACY

The accuracy of an analytical procedure expresses the closeness of agreement between the value which, is accepted either as a conventional true value or an accepted reference value and the value found.

This is sometimes termed trueness.

EXACTITUDE

L'exactitude d'une méthode analytique exprime le degré de concordance entre la valeur prise comme valeur de référence et la valeur retrouvée par la méthode analysée.

Cela est parfois nommé justesse.

PRECISION

The precision of an analytical procedure expresses the closeness of agreement (degree of scatter) between a series of measurements obtained from multiple sampling of the same homogeneous sample under the prescribed conditions.

Precision should be investigated using homogeneous, authentic samples.

The precision of an analytical procedure is usually expressed as the variance, standard deviation or coefficient of variation of a series of measurements.

Precision may be considered at three levels: repeatability, intermediate precision and reproducibility.

- Repeatability

Repeatability expresses the precision under the same operating conditions over a short interval of time. Repeatability is also termed intra-assay precision.

- Intermediate precision

Intermediate precision expresses within-laboratories variations: different days, different analysts, different equipment, etc.

- Reproducibility

Reproducibility expresses the precision between laboratories (collaborative studies, usually applied to standardization of methodology).

PRÉCISION

La précision d'une méthode analytique exprime le degré de concordance (degré de dispersion) entre une série de mesures obtenues à partir d'échantillons multiples homogènes réalisés dans les conditions définies.

La précision devrait être étudiée en utilisant des échantillons authentiques, homogènes.

La précision d'une méthode d'analyse est habituellement exprimée par la variance, l'écart type ou le coefficient de variation d'une série de mesures.

La précision peut être considérée à trois niveaux: répétabilité, la précision intermédiaire et la reproductibilité.

Répétabilité

La répétabilité exprime la précision dans les mêmes conditions d'exploitation sur un court intervalle de temps. La répétabilité est également nommée précision intra-essai.

Précision intermédiaire

La précision intermédiaire exprime variations au sein d'un laboratoire en fonction des jours différents, de manipulateurs différents et d'équipements différents.

Reproductibilité

La reproductibilité exprime la précision entre les laboratoires (intérêt dans des études collectives, généralement appliquée à la standardisation de la méthodologie).

ROBUSTNESS

The robustness of an analytical procedure is a measure of its capacity to remain unaffected by small, but deliberate variations in method parameters and provides an indication of its reliability during normal usage.

ROBUSTESSE

La robustesse d'une méthode analytique est une mesure de sa capacité à rester non perturbée par des petites mais délibérées variations dans les paramètres de méthode et fournit ainsi une indication de sa fiabilité durant une utilisation normale.

BIBLIOGRAPHIE

1. John P. A. Ioannidis. Why Most Published Research Findings Are False? PLoS Medicine August 2005 | Volume 2

2. George C.M. P.A. Ioannidis. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol. 2015 Jan;68(1):25-34.

3. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov.

4. Begley CG, Ellis L. Drug development: raise standards for preclinical research. Nature. 2012;483:531–533

5. Peers IS, Ceuppens PR, Harbron C. In search of preclinical robustness. Nat Rev Drug Discov. 2012;11:733–734.

6 Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet. 2009;374:86–89.

7. Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, Schulz KF, Tibshirani R. Increasing value and reducing waste in research design, conduct, and analysis. Lancet. 2014;383:166–175.

8. Tsilidis KK, Panagiotou OA, Sena ES, Aretouli E, Evangelou E, Howells DW, Al-Shahi Salman R, Macleod MR, Ioannidis JP. Evaluation of excess significance bias in animal studies of neurological diseases. PLoS Biol. 2013;11:e1001609.

9. Perrin S. Make mouse studies work. Nature. 2014;507:423–425.

10. Ioannidis JP, Allison DB, Ball CA, et al. Repeatability of published microarray gene expression analyses. Nat Genet. 2009; 41:149–155.

11. Sterne JA, Davey Smith G (2001) Sifting the evidence—What's wrong with signifi cance tests. BMJ 322: 226–231.

12. Wacholder S, Chanock S, Garcia-Closas M, El ghormli L, Rothman N (2004) Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. J Natl Cancer Inst 96: 434–442.

13. Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405: 847–856.

14. Bossuyt PM et als.(2003) « Towards Complete and Accurate Reporting of Studies of Diagnostic Accuracy: The STARD Initiative ». Clin Radiol. 2003 Aug;58(8):575-80. Review.

15. Fermanian J : Mesure de l'agrément entre deux observateurs, cas qualitatif. Rev Epidemiol Sante Publique 1984 ; 32 : 140-7.

16. Van Stralen K.J., Jager K.J., Zoccali C., Dekker F.W. Agreement between methods. Kidney Int 2008; 74 (9): 1116-1120.

17. Journois D : Comparaison de deux variables : l'approche graphique (méthode de Bland et Altman). Rev Mal Respir 2004 ; 21 : 127-30.

18. C. Fuhrman, C. Chouaïd. Concordance de deux variables : les approches numériques. Revue des Maladies Respiratoires. Vol 21, N° 1 - février 2004 pp. 123-125

19. Nicolas Rognant, Justine Bacchetta, Laurent Juillard. Comparaison des méthodes d'estimation d'un paramètre quantitatif : évaluation de la concordance. Néphrologie ; Thérapeutique Volume 9, numéro 2 pages 92-97 (avril 2013)

20. Bland J.M., Altman D.G. Statistical methods for assessing agreement between two methods of clinical measurement Lancet 1986; 1 (8476) : 307-310

21. Cicchetti DV, Feinstein AR : High agreement but low kappa : II. Resolving the paradoxes. J Clin Epidemiol 1990 ; 43 : 551-8.

22. Feinstein AR, Cicchetti DV : High agreement but low kappa : I. The problems of two paradoxes. J Clin Epidemiol 1990 ; 43 : 543-9.

23. Koch, Gary G. (1982). "Intraclass correlation coefficient". In Samuel Kotz and Norman L. Johnson. Encyclopedia of Statistical Sciences 4. New York: John Wiley & Sons. pp. 213–217.

24. P. E. Shrout & Joseph L. Fleiss (1979). "Intraclass Correlations: Uses in Assessing Rater Reliability". Psychological Bulletin 86 (2): 420–428. doi:10.1037/0033-2909.86.2.420. PMID 18839484.

25. Kottner J et als. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011 Jan;64(1):96-106.

26. Minutoli F, Di Bella G, Mazzeo A, Donato R, Russo M, Scribano E, Baldari S. Comparison between (99m)Tc-diphosphonate imaging and MRI with late gadolinium enhancement in evaluating cardiac involvement in patients with transthyretin familial amyloid polyneuropathy. AJR Am J Roentgenol. 2013 Mar;200(3):W256-65.

27. Krepkin K, Won E, Ramaswamy K, Triolo M, Stiffelma M, Rusinek H, Chandarana H. Dynamic Contrast-Enhanced MR renography for Renal Function evaluation in Ureteropelvic junction obstruction: Feasibility Study. AJR Am J Roentgenol. 2014 Apr; 202(4):778-83.

28. Kundu P, Lata S, Sharma P, Singh H, Malhotra A, Bal C. Prospective evaluation of (68) Ga-DOTANOC PET-CT in differentiated thyroid cancer patients with raised thyroglobulin and negative (131)I-whole body scan: comparison with (18)F-FDG PET-CT. Eur J Nucl Med Mol Imaging. 2014 Jul;41(7):1354-62.

29. Franz Buchegger et al. First imaging results of an intra-individual comparison of 11C-acetate and 18F-fluorocholine PET/CT in patients with prostate cancer at early biochemical first or second relapse after prostatectomy or radiotherapy. European journal of nuclear medicine and molecular imaging. 10/2013; 41(1).

30. Kubiessa K. Initial clinical results of simultaneous 18F-FDG PET/MRI in comparison to 18F-FDG PET/CT in patients with head and neck cancer. Eur J Nucl Med Mol Imaging. 2014 Apr;41(4):639-48.

31. Irshad A1, Ackerman SJ, Spicer K, Baker NL, Campbell A, Anis M, Shazly M. Ultrasound evaluation of gallbladder dyskinesia: comparison of scintigraphy and dynamic 3D and 4D ultrasound techniques. AJR Am J Roentgenol. 2011 Nov;197(5):1103-10.

32. Omoumi P1, Bafort AC, Dubuc JE, Malghem J, Vande Berg BC, Lecouvet FE. Evaluation of rotator cuff tendon tears: comparison of multidetector CT arthrography and 1.5-T MR arthrography. Radiology. 2012 Sep;264(3):812-22.

33. Nielsen KR, Chakera AH, Hesse B, Scolyer RA, Stretch JF, Thompson JF, Nielsen MB, Uren RF, Oturai PS. The diagnostic value of adding dynamic scintigraphy to standard

delayed planar imaging for sentinel node identification in melanoma patients. Eur J Nucl Med Mol Imaging. 2011 Nov;38(11):1999-2004.

34. Revel MP, Faivre JB, Letourneau T, Henon H, Leys D, Delannoy-Deken V, Remy-Jardin M, Remy J. Patent foramen ovale: detection with nongated multidetector CT. Radiology. 2008 Oct; 249(1):338-45.

35. Drubach LA, Zurakowski D, Palmer EL 3rd, Tracy DA, Lee EY. Utility of salivagram in pulmonary aspiration in pediatric patients: comparison of salivagram and chest radiography. AJR Am J Roentgenol. 2013 Feb; 200(2):437-41.

36. Chung JH, Little BP, Forssen AV, Yong J, Nambu A, Kazlouski D, Puderbach M, Biederer J, Lynch DA. Proton MRI in the evaluation of pulmonary sarcoidosis: comparison to chest CT. Eur J Radiol. 2013 Dec;82(12):2378-85.

37. Lei Wang, Chaowu Yan, Shihua Zhao, Wei Fang. Comparison of 99mTc-MIBI SPECT/18F-FDG PET Imaging and Cardiac Magnetic Resonance Imaging in Patients With idiopathic Dilated Cardiomyopathy. Clin Nucl Med 2012;37: 1163-1169

38. Valentina Ambrosini et al. 68Ga-DOTA-NOC PET/CT in comparison with CT for the detection of bone metastasis in patients with neuroendocrine tumours. Eur J Nucl Med Mol Imaging (2010) 37:722–727

39. Zink S, Ohki SK, Stein B, Zambuto DA, Rosenberg RJ, Choi JJ. Noninvasive evaluation of active lower gastrointestinal bleeding: comparison between contrast-enhanced MDCT and 99mTc-labeled RBC scintigraphy. AJR Am J Roentgenol. 2008 Oct;191(4):1107-14.

40. Acid S, Le Corroller T, Aswad R, Pauly V, Champsaur P. Preoperative imaging of anterior shoulder instability: diagnostic effectiveness of MDCT arthrography and comparison with MR arthrography and arthroscopy. AJR Am J Roentgenol. 2012 Mar; 198(3):661-7.

41. Duan Y, Wang X, Yang X, Wu D, Cheng Z, Wu L. Diagnostic efficiency of low-dose CT angiography compared with conventional angiography in peripheral arterial occlusions. AJR Am J Roentgenol. 2013 Dec;201(6):W906-14.

42. Rui Wanga et al. Low dose prospective ECG-gated delayed enhanced dual-source computed tomography in reperfused acute myocardial infarction comparison with cardiac

magnetic resonance. European Journal of Radiology 80 (2011) 326–330

43. Vivier PH, Sallem A, Beurdeley M, Lim RP, Leroux J, Caudron J, Coudray C, Liard A, Michelet I, Dacher JN. MRI and suspected acute pyelonephritis in children: comparison of diffusion-weighted imaging with gadolinium-enhanced T1-weighted imaging. Eur Radiol. 2014 Jan;24(1):19-25.

44. Fleiss J.L. Inference about weighted Kappa in the non-null case, Appl. Psychol. Meas., 1978, 1, 113-117.

45. Zink SI, Ohki SK, Stein B, Zambuto DA, Rosenberg RJ, Choi JJ, Tubbs DS. Noninvasive evaluation of active lower gastrointestinal bleeding: comparison between contrast-enhanced MDCT and 99mTc-labeled RBC scintigraphy. AJR Am J Roentgenol. 2008 Oct; 191(4):1107-14.

46. Femoral and Tibial Torsion measurements With 3D Models based on Low-Dose Biplanar radiographs in Comparison With standard CT Measurements. AJR 2012; 199:W607–W612

47. Mohamed E. Abou-El-Ghar, Ahmed El-Assmy, Huda Refaie, Tarek El-Diasty. Bladder Cancer: Diagnosis with Diffusion-weighted MR Imaging in Patients with Gross Hematuria. Radiology: Volume 251: Number 2—May 2009

48. Luca Saba, Giorgio Mallarini. A comparison between NASCET and ECST methods in the study of carotids: Evaluation using Multi-Detector-Row CT angiography. European Journal of Radiology 76 (2010) 42–47

49. Fuster D, Solà O, Soriano A, Monegal A, Setoain X, Tomás X, Garcia S, Mensa J, Rubello D, Pons F.A prospective study comparing whole-body FDG PET/CT to combined planar bone scan with 67Ga SPECT/CT in the Diagnosis of Spondylodiskitis. Clin Nucl Med. 2012 Sep;37(9):827-32.

50. Michielsen K et al. Whole-body MRI with diffusion-weighted sequence for staging of patients with suspected ovarian cancer: a clinical feasibility study in comparison to CT and FDG-PET/CT. Eur Radiol. 2014 Apr;24(4):889-901

51. Buhk JH, Kallenberg K, Mohr A, Dechent P, Knauth M. Evaluation of angiographic computed tomography in the follow-up after endovascular treatment of cerebral aneurysms--a comparative study with DSA and TOF-MRA. Eur Radiol. 2009 Feb; 19(2):430-6.

52. Beatriz Rodríguez-Vigil et al. « PET/CT in Lymphoma: Prospective Study of Enhanced Full-Dose PET/CT Versus Unenhanced Low-Dose PET/CT». J Nucl Med. 2006;47:1643-1648.

53. Sona Balogova et al. « Whole-Body 18F-Fluorocholine (FCH) PET/CT and MRI of the Spine for Monitoring Patients With Castration-Resistant prostate Cancer Metastatic to Bone ». Clin Nucl Med 2014;39: 951Y959

Titre en français
Évaluation de la concordance dans les études d'imagerie diagnostique :
Une étude de la qualité des données publiées

Résumé (français) :

OBJECTIF : Le problème de reproductibilité des résultats concerne les études d'imagerie diagnostique, en particulier celle de concordance. Une méthodologie rigoureuse et clairement définie est nécessaire pour planifier ce type d'étude. Ce travail propose d'analyser la qualité méthodologique d'un échantillon représentatif de travaux d'imagerie médicale à visée diagnostique et de comparer les résultats obtenus par un médecin junior et deux médecins seniors.

MATERIELS ET METHODES : Nous avons sélectionné 8 revues d'imagerie médicale (4 de radiologie et 4 de médecine nucléaire) de bon niveau publiées durant les 10 dernières années. Nous avons effectué une recherche exhaustive initiale par mots clés dans la base de données MEDLINE (via PubMed) et obtenu 372 articles. Plusieurs analyses successives (par lecture du résumé puis du texte) ont permis de constituer un échantillon de 80 articles. Un formulaire standardisé d'extraction de données a été généré qui comprenait 29 items relatifs à la qualité de l'étude notamment sur les biais méthodologiques. Tous les articles sélectionnés ont été analysés par trois relecteurs indépendants (1 interne et 2 médecins expérimentés). La concordance entre junior et senior est comparée de façon qualitative.

RESULTAT : Sur 80 articles analysés en intégralité, nous avons retenu 12 items qui méritent d'être discutés. Notamment plus de 90 % des articles avaient quatre éléments manquants: «mentionner si les évaluateurs appartiennent au même centre », «donner une justification de la taille de l'échantillon », «notification des écarts au protocole » et «discussion de la pertinence clinique de l'interprétation ». Il existe une bonne concordance entre junior et sénior.

CONCLUSION : Plusieurs points pourraient être améliorés dans la présentation des résultats des études d''imagerie diagnostique afin d'assurer la fiabilité de leurs conclusions.

Mots clés (français) : Concordance, Reproductibilité, Qualité méthodologique, Imagerie diagnostique, Revue systématique

Titre en anglais : Evaluation of agreement in diagnostic imaging trials:
a review of the quality of published data

Abstract (English) :
Background: The relevance of imaging procedures may notably vary according to clinical setting which calls for valid evaluation methods applicable to diverse clinical settings. The problem of reproducibility of the results concerns also the diagnostic imaging studies. Key issues center on substitution potential for novel radiopharmaceuticals, or easier to use medical devices vs. reference diagnostic tools. A clearly defined step wise methodology is required to plan the study and adequate statistical tests are required to assess the agreement between both methods.
The objectives of this study is to assess the quality and the relevance of statistical analyses in reports of recently published diagnostic medical imaging trials and to compare the results obtained by a junior doctor and two senior doctors..

Methods: We conducted a systematic review of literature in the MEDLINE database (via PubMed) to identify reports from 8 core journals (4 of nuclear medicine and 4 of radiology) with high impact factor, indexed over the past ten years. We have screened all potentially relevant articles with key words for an exhaustive search and assessed them to have final eligible articles which had been all full-text analyzed. A standardized data-extraction form was generated which comprised 29 items related to the study quality. We tried to identify all sources of methodological biases and missing data that could influence the interpretation of results. All selected articles were analyzed by 3 independent reviewers (1 junior and 2 seniors). Relevant items will be presented.

Results: Eighty articles have been full-text analyzed. Twelve items are deserved being discussed. Particularly five of them because more than 90% of publications had these 5 items missing: to have “agreement” or “comparison” in their key words, to mention if the assessors belong to the same center or not, to give a justification of the sample size, to report if there were protocol deviations, and to discuss the clinical relevance of interpretation. There is a good agreement between junior and senior.

Conclusions: Our findings show that several points could be improved in the reporting of diagnostic imaging trials to ensure the reliability of the conclusions.

Keywords (English) : agreement, reproducibility, systematic review, methodological quality, diagnostic imaging

Université Paris Descartes
Faculté de Médecine Paris Descartes
15, rue de l’Ecole de Médecine
75270 Paris cedex 06